Application of Principal Component Analysis to Galaxy Spectral Energy Distributions

KOONKOR, SUTTIKOON

Application of Principal Component Analysis to Galaxy Spectral Energy Distributions

Suttikoon Koonkor

A thesis submitted to Durham University
in accordance with the regulations for
admittance to MSc. by Research

Institute for Computational Cosmology
The University of Durham
United Kingdom

December 2020

Application of Principal Component Analysis to Galaxy Spectral Energy Distributions

Suttikoon Koonkor

Abstract

Galaxy spectra are a useful diagnostic tool that can be used to reveal the intrinsic properties of galaxies, such as their star formation rate and stellar mass, along with the conditions in the interstellar medium. Generally the computation of the full galaxy spectra within galaxy formation and evolution models tends to be very time consuming and memory inefficient, so the calculation of spectra is typically only done in post-processing for a subset of model galaxies (e.g. Trayford et al. 2017, Cowley et al. 2018). Upcoming surveys will measure tens of millions of spectra (e.g., Euclid (Laureijs et al. 2011) and the Dark Energy spectroscopic Instrument (DESI; Levi et al. 2019)). To exploit these data, theoretical models need to be able to predict spectra to connect more closely with these surveys. In this thesis, we aim to reduce the computational expense when calculating galaxy spectra by applying principal component analysis (PCA) to the spectral energy distributions of simple stellar populations (SSPs). We consider different star formation histories and different matallicities. As a result, we find that the dimensionality of the SSP spectra can be reduced by a factor of $\sim$50 whilst there is only a small loss in accuracy ($\sim$1 - 5%) of the reconstructed spectra. Moreover, we find that this loss in accuracy is negligible when computing broadband magnitudes ($\ll 1\%$). Our results suggest that this calculation method may be a plausible way to predict spectra for all the galaxies in the output of a semi-analytical model covering a cosmological volume (e.g. *GALFORM*; Cole et al. 2000).

Supervisors: Prof. Carlton Baugh and Dr. Peder Norberg

Acknowledgements

I would like to sincerely thank my supervisor, Prof. Carlton Baugh, for the endless encouragement, guidance, and valuable time he has given me throughout this research. I am also very grateful to Dr. Peder Norbergfor sharing his ideas and suggestions of computing technique. The opportunity to work with them is one of the greatest opportunities I have ever been given.

Moreover, I would also like to thank Dorothy Jenkins and my friends, Bingchao Wang, Xiao Jin, and Maura Ramirez-quezada for their mental supports during the pandemic lockdowns.

I acknowledge the financial, academic, and technical support from the Royal Thai Government Scholarship.

At last but not least, I thank my family from across the sea for their love, support, and belief.

Contents

Declaration

The work in this thesis is based on research carried out between 2019 and 2020 while the author was a research student under the supervision of Prof. Carlton Baugh and Dr. Peder Norberg in the Institute for Computational Cosmology, Department of Physics, University of Durham, England. No part of this thesis has been submitted elsewhere for any other degree or qualification.

List of Figures

List of Tables

List of Acronyms

AGB Asymptotic Giant Branch

AGN Active Galactic Nuclei

CDM Cold Dark Matter

CMB Cosmic Microwave Background

CSP Composite Stellar Population

FSPS Flexible Stellar Population Synthesis Fortran library

IMF Initial Mass Function

NN Neural Network

PCA Principal Component Analysis

SED Spectral Energy Distribution

SFH Star Formation History

SFR Star Formation Rate

SPS Stellar Population Synthesis

SSP Simple Stellar Population

Chapter 1

Introduction

Galaxy spectra are a useful diagnostic tool that can be used to reveal the intrinsic properties of galaxies, such as their star formation rate and stellar mass, along with the conditions in the interstellar medium. Following on from the Two-degree-Field Galaxy Redshift Survey (2dfGRS; Colless et al. 2001) and the Sloan Digital Sky Survey (SDSS; York et al. 2000), upcoming wide field surveys will measure tens of millions of spectra (e.g. Euclid: Laureijs et al. 2011; The Dark Energy Spectroscopic Instrument (DESI): Levi et al. 2019). To be able to connect more closely with these surveys, and to exploit the wealth of information contained in these observations, theoretical models need to be able to predict spectra. This will allow the model galaxies to be selected in as similar a way as possible to the selection of the observed galaxies, allowing us to build more realistic mock catalogues. Currently, galaxy formation models typically do not predict spectra directly. They predict the inputs needed to calculate the full spectral energy distribution, such as a galaxy's star formation history and its chemical evolution. However, the computation of spectra tends to be very time consuming and is only done for a subset of the model galaxies in post-processing (e.g. Trayford et al. 2017; Cowley et al. 2018). The objective of this thesis is to explore a fast way to add spectra to the model predictions for all galaxies as the model is running.

In this Introduction, we explain why the spectral energy distribution of a galaxy

is an interesting property to output, and we give a brief overview of galaxy formation models. Our framework can be implemented into these models; however, to speed up the development of the code, rather than run a full galaxy formation model we have used a simple parametric form for the star formation history of a galaxy and have made simple assumptions about how the metallicity of the star forming gas changes with time. The implementation of the code into a full model of galaxy formation is left for future work.

1.1 The spectral energy distribution

The spectral energy distribution of a galaxy records the flux emitted as a function of wavelength. Different components of the galaxy contribute: i) stars, ii) clouds of hydrogen ionised by energetic photons produced by massive young stars, called HII regions, iii) dust within the galaxy, iv) accretion on to a central supermassive black hole.

Fig. 1.1 shows the UV-to-NIR spectra of different types of galaxies. Moving from ellipticals (E, top) to late-type spirals (Sa-Sc; lower spectra), the continuum becomes bluer, with more photons emitted at short wavelengths, and the emission lines becomes systematically stronger. This sequence is approximately one of increasing star formation activity as we move down from the top. For early-type galaxies, which lack hot, massive young stars because they have little or no recent star formation, most of the light emerges at long wavelengths and the spectrum shows a small amount of light at wavelengths shorter than 4000Å and there are no emission lines. On the other hand, late-type galaxies and starbursts, which do have ongoing star formation emit more light in the blue and near-ultraviolet; this light is dominated by hot, massive young stars which have short lifetimes. Because of this the interstellar medium also gets heated and is ionized by the Lyman continuum photons giving rise to strong emission lines (see Byler et al. 2018). The 4000Å feature arises in old stellar populations due to the CaII (K, H) absorption features

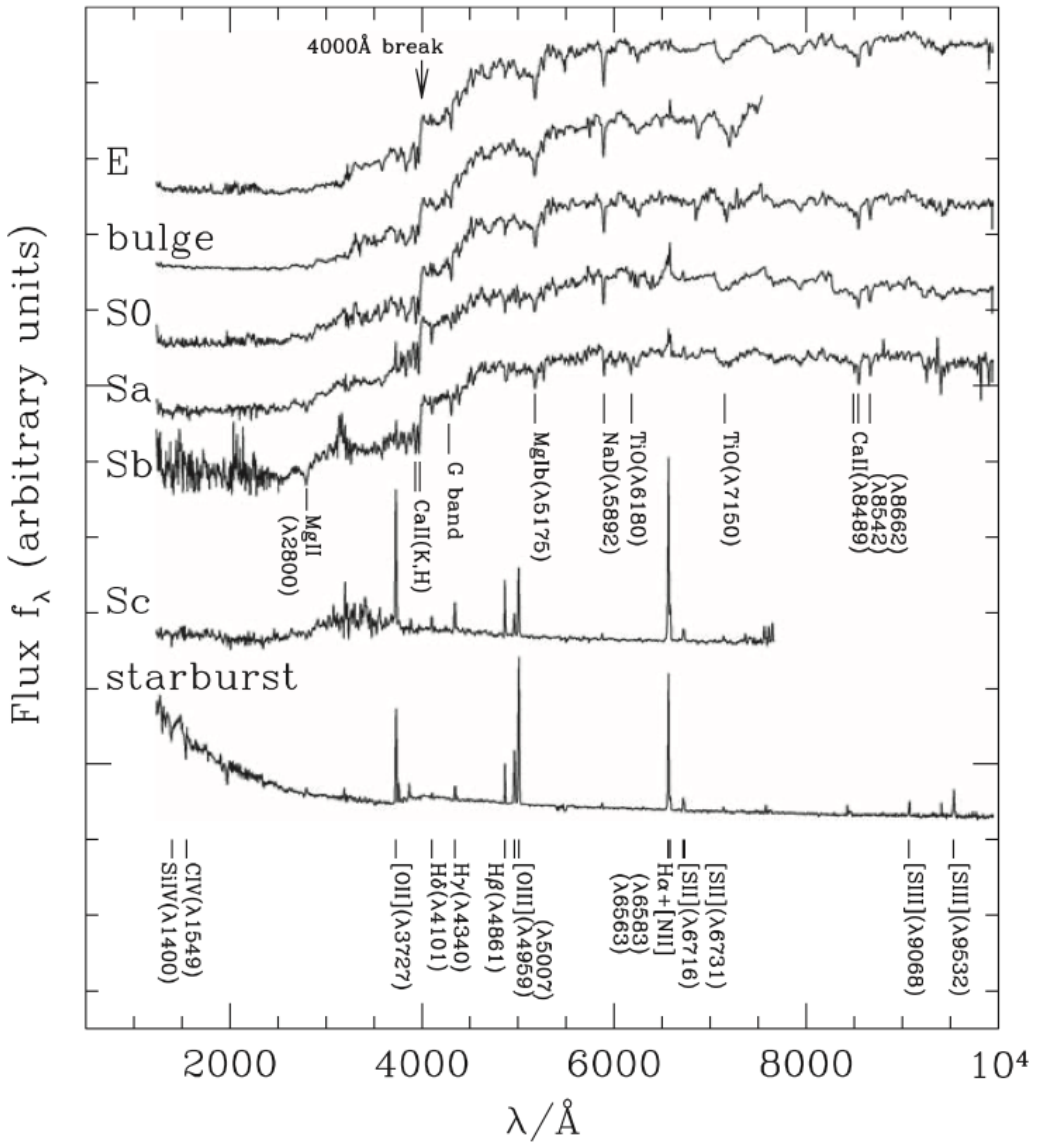


Figure 1.1: The UV-to-NIR spectra of different galaxy types from ellipticals to late-spirals and starburst. The emission and absorption lines (with the associated wavelengths) are shown and labeled by the vertical lines. The spectra are offset in amplitude for clarity. See text for details. Figure taken from Mo et al. 2010.

and other absorption lines. This feature is apparent as a drop in the Elliptical spectrum continuum, which becomes less pronounced moving down the spectra in the figure, as the stellar populations become progressively younger.

The stellar spectrum is attenuated by the dust grains, with the radiation that is absorbed being re-radiated at longer wavelengths. We do not consider the dust emission spectrum further in this thesis. We also do not consider emission lines

in the galaxy spectra further. Finally, we do not consider the contribution to a galaxy's SED from an nuclear activity. Instead we focus on the simplest prediction made by stellar population synthesis models (see Chapter 2); the stellar emission. We will consider the effects of the age and metallicity of the stellar population on the appearance of the spectrum in the next chapter.

1.2 An Overview of the Theory of Galaxy Formation

Whilst this thesis is not about the physics of galaxy formation, and we do not implement our code into a physical model at this point, here, for completeness we give a brief overview of galaxy formation modelling.

Modern galaxy formation theory is based on the hierarchical structure formation paradigm, in which small fluctuations in density, seeded during inflation, are amplified by gravity and grow into galaxies and groups and clusters of galaxies. The standard cosmological model, Λ cold dark matter (ΛCDM), is well constrained and supported by many observations including the temperature fluctuations of the cosmic microwave background (CMB) radiation (e.g. Komatsu et al. 2011; Planck Collaboration et al. 2018), the magnitude-redshift relation of Type Ia supernovae (e.g. Kowalski et al. 2008), and the large-scale structure of the Universe as measured by spectrocopic galaxy surveys of large scale structure (e.g. Cole et al. 2005; Eisenstein et al. 2005; Percival et al. 2007a; Reid et al. 2010)*. The standard ΛCDM universe contains two forms of energy-density. The first is dark energy that makes up the highest portion of the universe, accounting for about 68 percent of the total energy density today. And the rest is matter (dark matter (DM) and baryonic matter). The dark matter is referred to as non-relativistic (cold) collisionless particles that mainly interact through gravitation. The DM makes up 27

*We note that the model has been challenged by observations on small scales, such as the abundance and structure of satellite galaxies. Various solutions have been proposed, which include considering the impact of baryonic physics (see Weinberg et al. 2015 for a review).

percent of the universe. The remaining 5 percent is baryonic matter, namely the atoms that produce all of the light we can observe in the universe.

The basic theory behind how galaxies form and evolve in the hierarchical structure formation paradigm has been well established for many decades (Rees and Ostriker 1977; White and Rees 1978; White and Frenk 1991). Current models of galaxy formation typically follow the following processes: **(i) Gravity** - Gravity plays an important role in constructing the foundation for galaxy formation through the formation and merging of dark matter halos; **(ii) Hydrodynamics and Thermal evolution** - When gas and dark matter collapse in an over-dense region, the entropy and temperature of the gas can be increased by strong shocks. Then the formation of galactic disks is determined by how efficiently the gas can cool and radiate away the thermal energy; **(iii) Star formation** - in galaxy disks and in starbursts; **(iv) Feedback** - includes the effects due to supernovae, active galactic nuclei (AGNs), photo-ionization of the intergalactic medium (IGM); **(v) Galaxy mergers** - that can trigger starbursts and lead to the formation of spheroids from the effect of dynamical friction; spheroids can also form when disks become dynamically unstable, leading to bar instabilities that transfer material to the centre of the galaxy, possibly triggering a star burst – these events can be triggered by perturbations due to the presence of satellite galaxies; **(vi) Stellar population synthesis and chemical evolution**; this step allows us to make direct comparisons between models and observations by combining the predicted star formation histories and chemical evolution with a stellar population synthesis model.

There are two broad types of physical models of galaxy formation: gas simulations and semi-analytics. The first approach is the most explicit way to model galaxy formation by using numerical hydrodynamic techniques to solve the equations of gravity, hydrodynamics, and thermodynamics of particles and/or grid cells that represent dark matter, gas, and stars. On the other hand, the other technique, semi-analytic modelling (SAM), has been widely used to model galaxy formation.

This approach does not explicitly solve fundamental equations for particles or grid cells, but adopts a set of simplified equations instead and, in general, requires stronger approximations and assumptions to be made than in the case of gas simulations. In both cases, some physical processes remain "sub resolution" or "sub-grid" or simply, we do not know the correct equations to describe them – in these cases, both approach resort to "semi-analytic" approaches with parameters. As these two approaches make different assumptions and approximations, our focus here is not on the details of these models, but on trying to extend the functionality of the models by allowing a direct prediction of the SED. For a more detailed discussion of the two different approaches see Baugh (2006); Somerville and Davé (2015).

To make direct comparisons between models and observations, implementing a stellar population synthesis model directly into a galaxy formation model to compute the full spectrum can be computationally expensive when one attempts to calculate the SED for every single galaxy in the model. Currently, *GALFORM* pre-processes the output of the stellar population synthesis model to tabulate the mass-to-light ratios in a set of specified filters. The overall mass-to-light ratios in each band are computed for the composite stellar population, then multiplied by the stellar mass of the galaxy to obtain the magnitude in each band. If a magnitude is required in a different band, *GALFORM* has to be re-run. This approach has the advantage that only a few numbers have to be stored - the mass-to-light ratios in the bands. On the other hand, the full spectrum consists of more than a thousand wavelength bins. Therefore, in general, galaxy formation models do not provide galaxy spectra in their standard output. Instead, a subset of the model galaxies is selected in some way from the full model output to be a sample set to calculate the SEDs by using a post-processing calculation (e.g. Cowley et al. 2018; Trayford et al. 2017).

Even though a small sample of the models galaxy of $\sim 10^5$ galaxies out of millions galaxies is selected, the calculation time for the SED calculations by coupling the *GALFORM* with the full output of a stellar population synthesis model

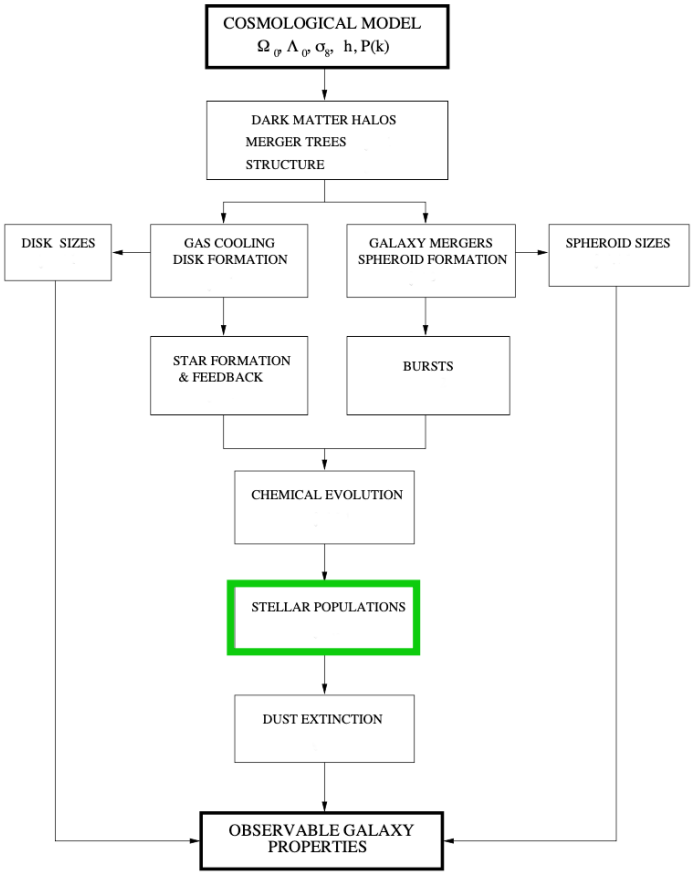


Figure 1.2: The schematic overview of *GALFORM*. The main objective of this study is to calculate galaxy SEDs, which is highlighted in green rectangle. Figure adapted from Cole et al. 2000.

GRASIL (Silva et al. 1998) is an significant computational overhead (Cowley et al. 2018). Therefore, the aim of this study is to reduce the computational expense of the calculation of spectra, which is highlighted in the green rectangle in Fig. 1.2.

1.3 Thesis Outline

In this thesis, we aim to reproduce galaxy spectra using PCA. The structure of the remainder of the thesis is as follows. In Chapter 2, we review how the spectrum of a model galaxy can be computed using a stellar population synthesis model. In Chapter 3, we give an overview of the mathematical background of PCA and describe how we can use it to reduce the dimensionality of the galaxy spectra. In

Chapter 4, we discuss the effect of the scaling the spectra on the performance of the PCA for spectral reconstruction (§4.1). Then we show the result of applying PCA to simple stellar populations with a fixed metallicity and different ages (§4.2). We apply the PCA to a 2D-grid of simple stellar populations with both age and metallicity varying in §4.3. In that Chapter we also show how the number of principal components is chosen based on the error in the SED reconstruction. In Chapter 5, we calculate the SED of a composite stellar population by using the combination of a parametric star formation history and the PCA spectra of the simple stellar population (§5.1). We also calculate the photometry of the composite stellar population obtained from the PCA approach and compare the result to the color-magnitude diagram observed for local galaxies to determine a realistic star formation history. Finally, Chapter 6 provides the conclusions of this study and gives suggestions for future work.

Chapter 2

Stellar Population Synthesis

In studying the formation and evolution of galaxies, stellar population synthesis (SPS) is the tool that allows us to build a spectrum for a model galaxy. It is a technique to model the spectophotometric properties of stellar populations using the understanding of the evolution of stars. In this chapter we will briefly discuss the history of the SPS technique, then provide an overview of the ingredients of these models (Fig. 2.1), show some basic features of SSPs (§2.1), and finally introduce the idea of composite stellar populations and how it is calculated (§2.2).

Stellar population synthesis (sometimes referred to as evolutionary population synthesis e.g., Maraston 1998) modeling has a rich history. It was pioneered by Tinsley (1968). This approach provides an analytical method to predict the spectrum of a stellar population by assuming a star formation history and an initial mass function that stars are produced with, combined with the evolution of a star at different stages on the Hertzprung-Russell (HR)-diagram, which is governed by its mass. The population synthesis technique was developed substantially throughout the 1980s and 1990s (e.g., Tinsley and Gunn 1976; Bruzual A. 1983; Bruzual A. and Charlot 1993; Worthey 1994; Leitherer et al. 1999).

Nowadays, there are several popular SPS models available e.g., Silva et al. 1998, *GRASIL*; Bruzual and Charlot 2003, *BC03*; Maraston 2005, *Ma05*; Conroy and Gunn 2010, *FSPS*. We do not aim to compare the differences between SPS

models in this study. For readers who are interested in this, see Conroy and Gunn (2010); Chen et al. (2010); Baldwin et al. (2018).

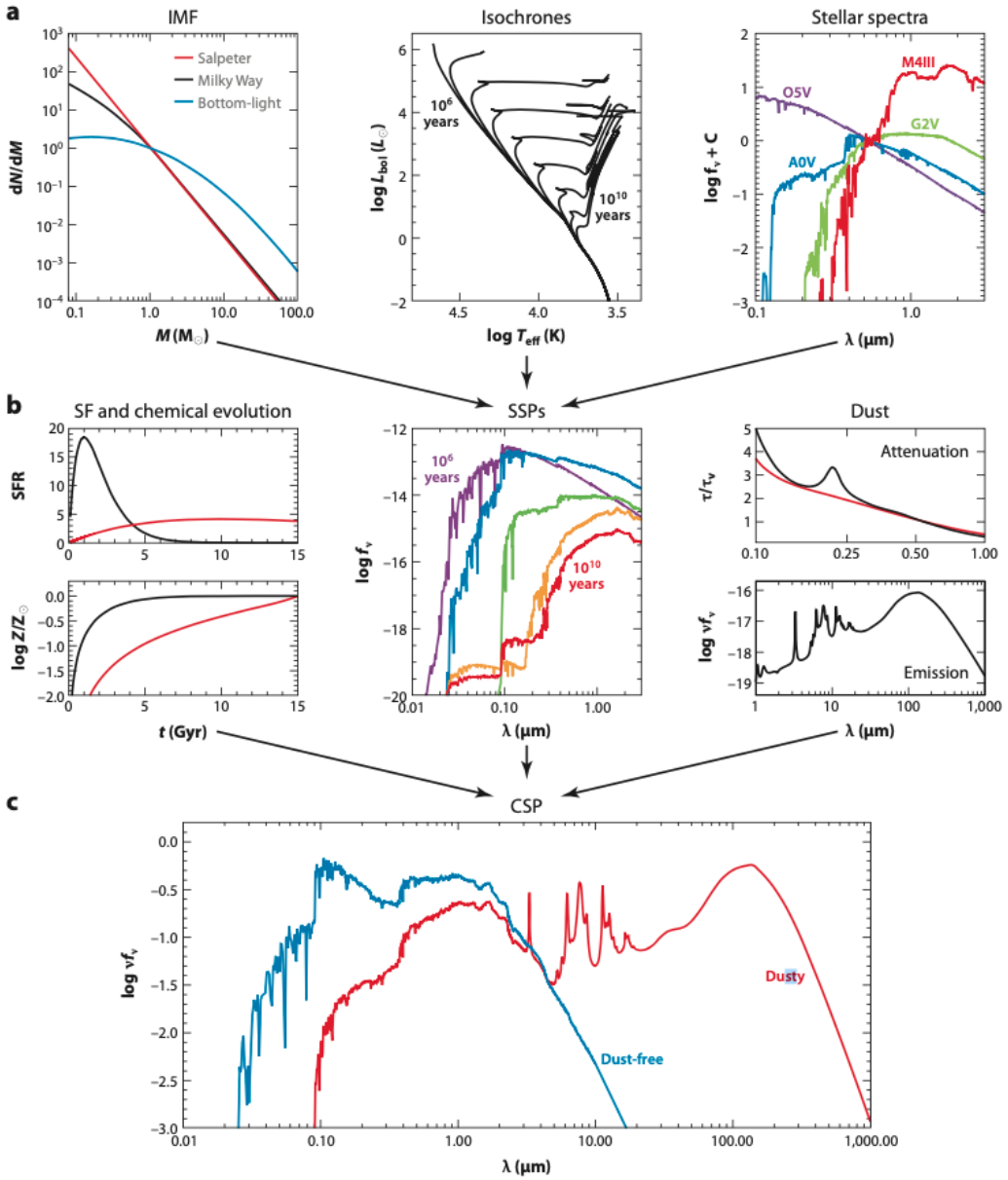
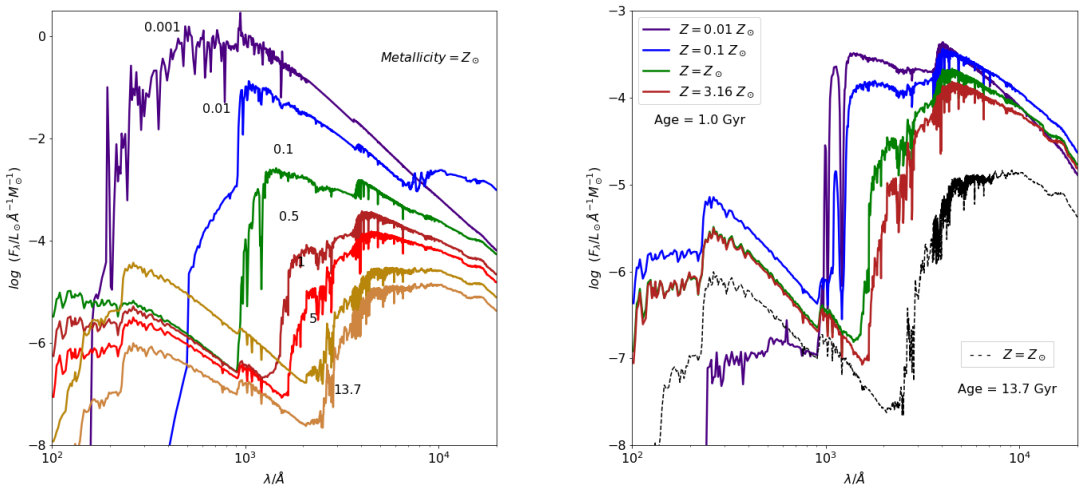


Figure 2.1: Overview of the stellar synthesis technique used in the FSPS: a) The top three panels show the main ingredients for constructing simple stellar populations (SSPs) including, from left to right, the stellar initial mass function (IMF), isochrones, and stellar spectra. b) The middle three plots show the key components of composite stellar populations (CSP), which include star formation histories (SFHs) and chemical evolution, SSPs, and dust attenuation and emission. c) The final result, the CSP showing stellar emission only (dust-free) and including the effects of dust (dusty). Figure taken from Conroy (2013).

2.1 The Simple Stellar Population



a) solar-metallicity SSPs at different ages

b) SSPs with different metallicities

Figure 2.2: a) The spectral energy distributions of a simple stellar population with solar metallicity, $Z_\odot$, at different ages (in Gyr). b) The spectral energy distributions of simple stellar populations with different metallicities at a fixed age of 1.0 Gyr (Solid line in colors) compared with the solar-metallicity simple stellar population at 13.7 Gyr (black dashed line). Both examples are predicted by using the FSPS model with a Kroupa IMF.

Fig. 2.2a shows the spectral energy distributions (SEDs) of simple stellar population (SSP) with a fixed solar metallicity at different ages and Fig. 2.2b shows the SEDs of simple stellar population at age 1.0 Gyr with different metallicities. The evolution of the simple populations can be understood by comparing the SEDs. At a very young age (< 1 Myr), the SED displays lots of emission in the UV because the blue main sequence stars that have high effective temperatures emit light strongly in the UV region. At about an age of few tens Myr, the most massive stars have evolved to become red supergiants. The death of the most massive stars causes a huge drop in the UV and a rise in the importance of the near-IR. The UV flux continues to drop during about 0.1 to 1 Gyr, but near-IR flux remains high because lower mass stars evolve to the asymptotic giant branch (AGB) phase. After that, red giant branch stars are the main contributors in the near-IR at a few Gyr. From the SEDs, there is remarkable rise in the UV at very old ages which is

the effect of low-mass stars evolving into post-AGB phase.

The evolution of a star also depends on its chemical composition; stars with higher metallicities evolve faster than low-metallicity stars. As shown in the right hand side of Fig. 2.2, a high-metallicity population at age 1 Gyr (red solid SED curve) looks similar to a very old age solar-metallicity population at age 13.7 Gyr (black dashed curve).

We have provided examples of the SED evolution of stellar populations above. To understand the evolution of a stellar population and to understand how is SED is computed, we consider the simplest case in which all stars in the population begin to evolve at the same time without any change in metallicity and no subsequent star formation, i.e. the simple stellar population. The key ingredients of constructing an SSP are an initial mass function (IMF), isochrones, and stellar spectral libraries (more details see Conroy, 2013, Section 2.1). These ingredients are listed below.

- **The Initial Mass Function (IMF):** gives the mass distribution of stars at their birth. The IMF of the Milky Way (MW) was established using the observational data by Salpeter (1955) to have the form of a power law

$$N(M)dM \propto M^{-x} \tag{2.1}$$

with $x = 2.35$ for masses greater than $\sim 0.5M_{\odot}$. Later works (e.g. Kroupa 2001 and Chabrier 2003) found that the Salpeter IMF overestimated the distribution of low-mass stars in the MW. A piece-wise power-law IMF was then proposed to lower the slope for low-mass stars with $x = 1.3$ at $M < 0.5M_{\odot}$ and $x = 2.3$ at $M \geq 0.5M_{\odot}$. Usually in galaxy formation models or stellar population synthesis models, the form of IMF is incorporated universally (independently of star formation history, morphology, metallicity, etc.). Even though the IMF affects the stellar mass-to-light ratio, the rate of luminosity evolution, and the shape of simple and composite stellar populations, our work does not aim to consider the effect of different forms of the IMF.

Throughout this study we use the Kroupa (2001) form. For a further discussion about the variation of the IMF in massive early-type galaxies, see the recent review by Smith (2020).

- **Isochrones:** The evolution of a star is almost determined by its initial (zero-age main sequence) mass and chemical composition. By probing the two most important properties of a stars which are the effective temperature, T_{eff}, and the luminosity, L, the evolution of the star can be represented in the $T_{eff} - L$ plane. As the T_{eff} and L are also related to the color (e.g. B-V) and absolute magnitude of a star, the evolutionary tracks of stars can be plotted on sa diagram called the Hertzsprung-Russell (HR) diagram. Figure 2.3 shows an illustration of the different evolutionary phases of a solar-mass star on the HR diagram.

  An isochrone then provides the same content as the evolutionary track, but instead of tracking stellar parameters as a function of age, it connects the parameters of different masses at the same age (see Figure 2.4a and Figure 2.4b). Figure 2.4 shows the stellar evolutionary tracks and the isochrones of stars with masses from 0.1 to 100 $M_{\odot}$ at different ages from 10^5 to 10^{10} years. Isochrones are usually constructed by calculating the stellar evolution from the hydrogen burning limit ($\approx 0.1 M_{\odot}$) to the maximum limit ($\approx 100 M_{\odot}$) depending on the model.

- **Stellar Spectral libraries:** To convert a model of stellar evolution into an observable SED, the stellar spectra of a specific metallity associated with the surface gravity and the effective temperature of stars in the population are required. There are 2 different approaches to obtain the stellar spectrum.

  The first is to use **Empirical Spectra**. An empirical library is based on observations of stars in the solar neighbourhood. An accurate spectrum is available for a star with the measured absolute magnitude, effective temperature and metallicity. Then the spectrum of a star with given a metallicity

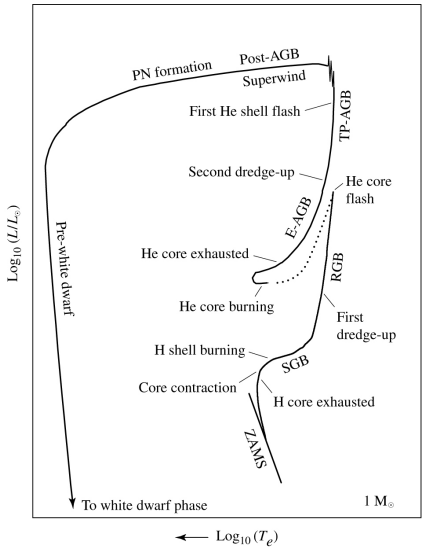
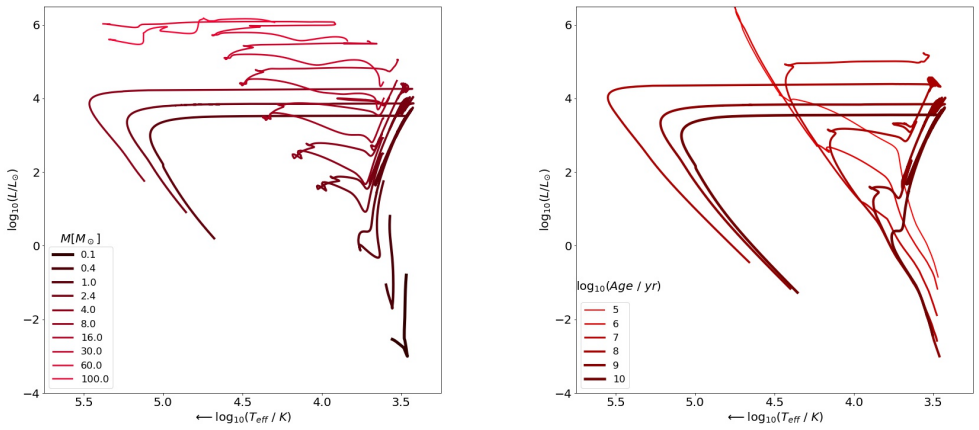


Figure 2.3: A schematic evolutionary track of a solar-mass star on the HR diagram. Figure taken from Carroll and Ostlie (1996).



a) Evolutionary Tracks b) Isochrones

Figure 2.4: a) The evolutionary tracks of solar-metallicity stars with different masses. b) The plot shows the isochrone of stars with the same physical properties as shown on the left, but rather than tracking stars of a common mass, the isochrone connects stars with the same age instead. The data used in these plots are taken from Choi et al. (2016).

and effective temperature can be calculated by interpolation. One of the first comprehensive observational spectral libraries was provided by Gunn and Stryker (1983). (For more details, e.g. current optical libraries see Yan and MaStar Team, 2017, Table 1.)

The second approach is to use **Theoretical Spectra**. The advantage of the

theoretical library is the broader coverage of parameter space and improved spectral resolution compared to the empirical one. A synthesis spectrum is calculated from the input atomic and molecular parameters and assumptions of the stellar atmosphere. For further discussion and a list of the libraries available see e.g. Conroy 2013, Section 2.1.3 and Mo et al. 2010, Section 10.3.1.

As we have provided the definition and an overview of the three SPS ingredients, we list the choice used in this study in Table 2.1 below.

SPS Ingredients	Model used in this study
IMF	Kroupa IMF: $x = 1.3$ at $M < 0.5M_{\odot}$ $x = 2.3$ at $M > 0.5M_{\odot}$ (Kroupa, 2001)
Isochrone	MIST: $-2.5 \leq \log[Z/Z_{\odot}] \leq 0.5$ with $Z_{\odot} = 0.0142$ $5 \leq \log(Age/yr) \leq 10.3$ $0.1 \leq M/M_{\odot} \leq 300$ (Choi et al., 2016)
Spectral Library	MILES empirical spectral library (Sánchez-Blázquez et al., 2006)

Table 2.1: The summary of the SPS model ingredients used in our study. For the coverage of the MILES library see Figure 2 of Conroy 2013.

The SED of a simple stellar population given its age and metallicity, $\mathcal{L}_{\lambda}^{SSP}(t, Z)$, can be constructed by combining these three ingredients as follows:

$$\mathcal{L}_{\lambda}^{SSP}(t, Z) = \int_{m_{lo}}^{m_{up}(t)} L_{\lambda}^{star}[T_{eff}(M, t, Z), L(M, t, Z)]\Phi(M)dM, \qquad (2.2)$$

where L_{λ}^{star} is a stellar spectrum from the stellar spectral library determined by the effective temperature (T_{eff}) and the bolometric luminosity (L) of a star with mass M and metallicity Z, $\Phi(M)$ is the IMF, and M is the initial stellar mass. The lower limit of integration, m_{lo} is generally referred to the hydrogen burning limit and the upper limit mass is more uncertain and typically take to be of the order $100M_{\odot}$.

2.2 The Composite Stellar Population

In reality, a galaxy (i.e. a population of stars mixing with cold gas, nebulae, AGNs etc.) is more complicated than the single-metallicity coeval stellar population that we have described in Section 2.1. Stars in a galaxy are produced at different times with the rate of star formation which is described by the star formation history (SFH).

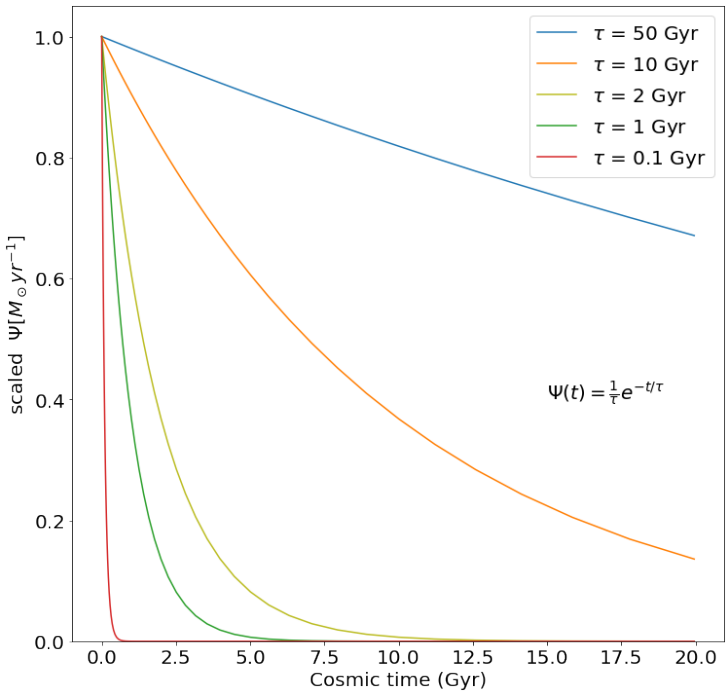


Figure 2.5: The star formation history of the tau model with different tau values (see Equation 2.5)

When stars die out, they leave behind stellar remnants. They produce winds ejecting mass and metals as they evolve. This can change the composition of the next generation stars. Moreover, a galaxy is mixed with stars and dust together. An observed SED of a real galaxy, therefore, must be more complex. To understand the stellar component of a galaxy SED, we consider composite stellar populations (CSPs) which differ from simple populations in three respects: (1) stars in a CSP have a range of ages given by the star formation history (SFH); (2) they contain stars with different metallicities described by their time-dependent metallicity dis-

tribution function $P(Z,t)$; and (3) the stars are produced in regions that contain dust (Conroy, 2013).

In Section 2.1, we have provided an overview of SSPs which are the building blocks for composite stellar populations. A composite population can be computed by combining each building block using the star formation history, chemical evolution, and dust in the following way:

$$\mathcal{L}_{CSP}(t) = \int_{t'=0}^{t} \int_{Z=0}^{Z_{max}} \left[SFR(t-t')P(Z,t-t')\mathcal{L}_{SSP}(t',Z)f_{dust,abs} + f_{dust,em}\right] dt'dZ, \tag{2.3}$$

where t' and Z are the integration variables referring to the population age and metallicity, respectively. The dust absorbs starlight, particularly at short wavelengths, and as a result the dust gets heated and can reradiate the energy at longer wavelengths (see the plots on the left in Fig. 2 from Cowley et al. 2018 for example). The model of dust absorption and emission can be added to the SPS as $f_{dust,abs}$ and $f_{dust,em}$ in Equation 2.3.

Despite the fact that the light propagating through the geometry of a galaxy is affected by dust which is mixed within the galaxy, we do not aim to take the account of dust in this study as we can deal with its effect separately, in post-processing. Moreover, we only consider simpler populations for which a single metallicity is assumed for the entire composite stellar population. Therefore Equation 2.3 becomes much more simple as

$$\mathcal{L}_{CSP}(t) = \int_{t'=0}^{t} SFR(t-t')\mathcal{L}_{SSP}(t',Z)dt'. \tag{2.4}$$

Even though the SFH obtained from a galaxy formation simulation can be very complicated (e.g. see Fig. 2 of Cowley et al. 2018), a simpler SFH are usually assumed in inferring the galaxy properties (e.g. Mitchell et al. 2013; Simha et al. 2014). One of the most popular form of SFH is the delayed exponential SFH,

$$SFR(t,\tau) = \frac{t}{\tau}e^{-t/\tau}, \tag{2.5}$$

where t is the time that has elapsed since the beginning of the SFH and τ is the characteristic e-folding timescale, which is a model parameter. This form of SFH includes an early rising SFR (linear term, t/τ) and a late-time decaying SFR (exponential term, $e^{-t/\tau}$) which are the natural consequence of galaxy evolution in a hierarchical Universe (e.g., high redshift galaxies see Maraston et al. 2010 and Papovich et al. 2011) and the scenario of a closed-box model (e.g., see Schmidt 1959), respectively.

Despite the apparently poor match between parametric SFHs and the SFH obtained from a model (see the examples plotted in Baugh 2006), we are not interested in the precise form of the SFH. We will use a SFH of the form of a tau model with different tau values for computing the composite stellar population (§5) in this work, as eventually this will be replaced by the one calculated by *GALFORM*. By using a parametric form for the SFH, the development of our PCA code is greatly sped up.

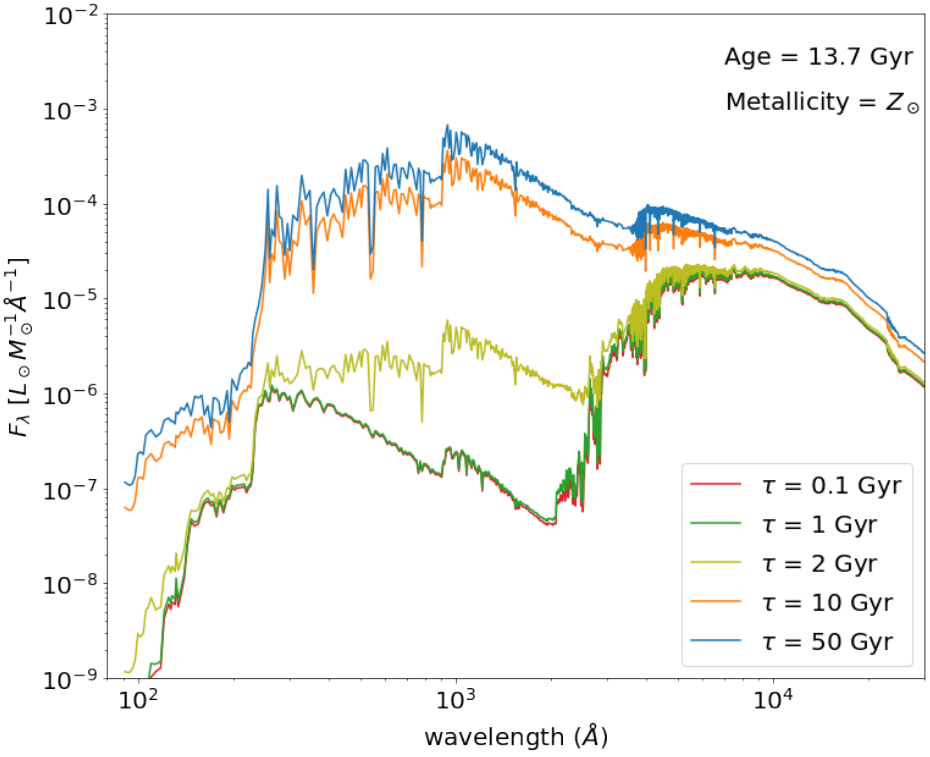


Figure 2.6: The SED of a solar-metallicity composite stellar populations viewed at an age of 13.7 Gyr, computed for different τ values, the same as those plotted in Fig. 2.5.

The synthesis model outputs a finite number of SSPs. SSPs can then be generated at output times that are not on the original grid using interpolation. We can change the form of Eq. 2.4 from an integration to a summation as

$$\mathcal{L}_{CSP}(t_{age}) = \sum_{i=1}^{n_{tage}} \mathcal{L}_{SSP,i}\ w_i, \tag{2.6}$$

where t_{age} is the age of the CSP, n_{tage} is the index of the oldest SSP (i.e. the index that corresponds to the age of the CSP) and w_i is the SFH weight of the SSP which is defined as $w_i = \int_{t_{i-1}}^{t_i} SFR(t_i - t')dt' / \int_0^{t_{age}} SFR(t_{age} - t')dt'$.

We found good agreement with the calculation made by FSPS when we apply 1600 bins in the CSP calculation. Fig 2.7 shows a comparison between a CSP computed directly from the FSPS code and our calculation. The sense of error when comparing the result from the calculation and the expectation is that a positive error refers to the overprediction whilst the underprediction is shown by a negative value and this sense will be applied throughout the thesis. Hence this number of age bins will be used in all CSP calculations in this study.

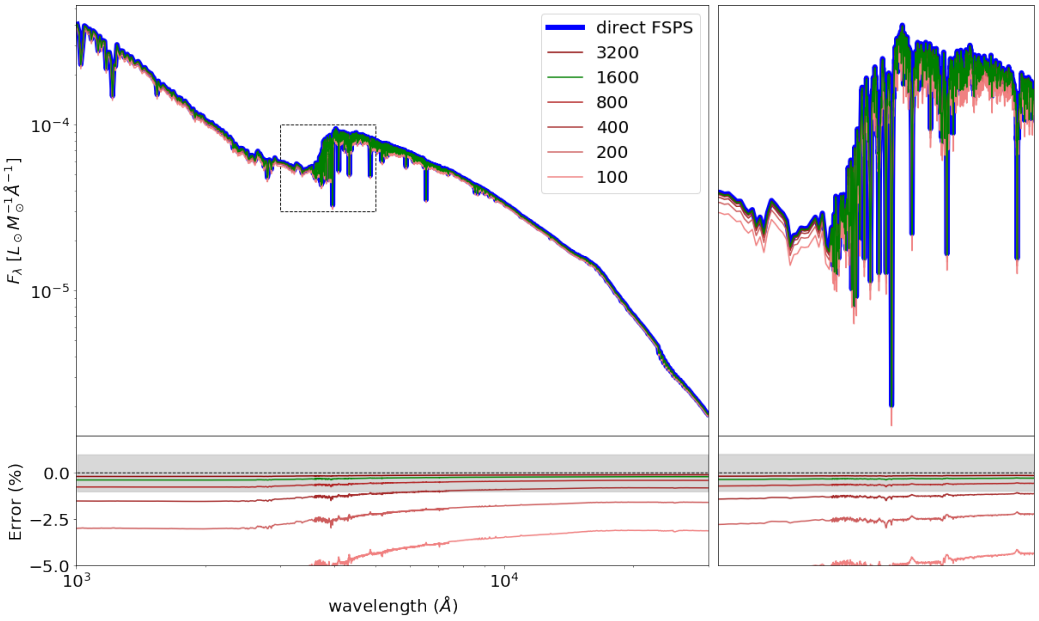


Figure 2.7: Top left: A comparison between the CSP calculation direct from FSPS code (blue line) and the calculation by using Equation 2.6 (red and green lines). Green line represents the CSP calculated by using 1600 age bins. Top right: The zoom-in spectrum on 3000 - 5000 Å spectrum shown as a rectangular area on the left plot. The bottom panels show the percentage error of the CSP spectra compared to the direct FSPS code.

Chapter 3

Principal Component Analysis

A galaxy spectral energy distribution (SED) typically consists of thousands of pieces of information when expressed as a function of wavelength. This can be condensed into a few numbers by measuring spectral features, such as the 4000 Angstrom break, or by sampling the spectrum using broad band filters, which can in turn be related to intrinsic galaxy properties (e.g. Kauffmann et al. 2003, Gallazzi et al. 2005). It is obviously complicated to make use all of the information contained in a galaxy SED as it consists of a vast amount of data. On the other hand, some important information stored in a galaxy SED may be lost when the full SED is replaced by "summary" statistics or measurements. In this chapter we will introduce a data compression technique called principal component analysis (hereafter, PCA) and provide a detailed mathematical overview. Finally we will show how we can reduce the dimensionality of galaxy SEDs by using PCA, whilst, at the same time, retaining the full information in the SED.

3.1 Principal Component Analysis

Principal component analysis (PCA) is a well known statistical technique that has been proven to be useful for dealing with high dimensional data in astronomy (see e.g. Connolly et al. 1995 and references therein) and it will be the main mathematical tool used throughout this study.

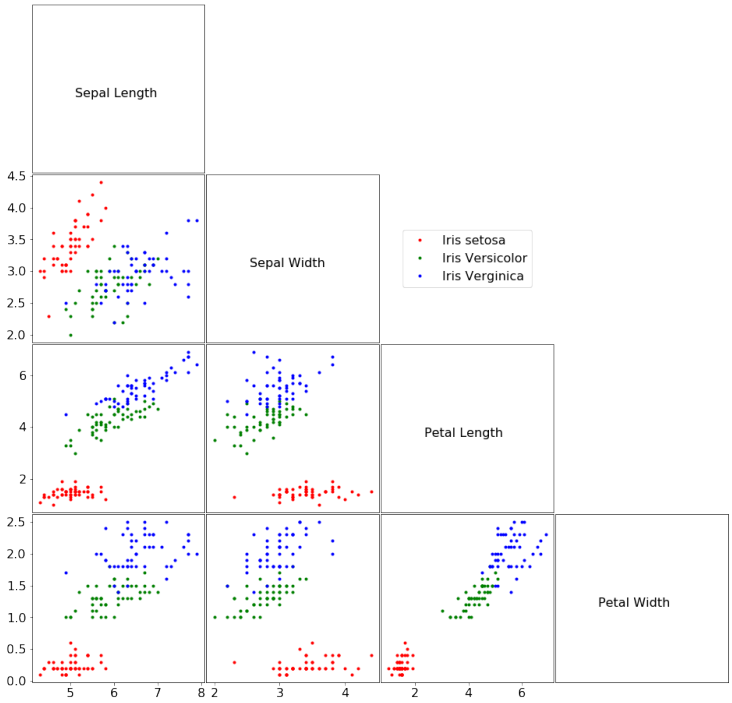


Figure 3.1: The scatterplot of the iris flower data set Fisher 1936 comparing measurements of different properties. The three different types of iris flowers including Setosa, Versicolor, and Verginica are shown as red, green, and blue dots, respectively. The data used in this plot are from Table 3.1.

PCA is an unsupervised technique used to reduce the dimensionality of data sets. It defines a new set of uncorrelated axes and reorders their importance according to the amount of variance along each of the new axes. Once the new set of axes is specified, the original data can be mapped onto it. We demonstrate the application of the PCA technique using a well known classification example called the Iris flower data set (Fisher, 1936). This catalogue of iris flowers contains 150 examples of three related iris flower species. Each entry is described by four characteristics (i.e. features) including sepal length, sepal width, petal length, and petal width as shown in Table 3.1.

Each sample in the original iris data set can be described as a vector in a 4 dimensional space as

$$\vec{x}_i = A_i\hat{e}_1 + B_i\hat{e}_2 + C_i\hat{e}_3 + D_i\hat{e}_4, \tag{3.1}$$

where A_i, B_i, C_i, and D_i represent, respectively, the values of measurements

	Sepal Length	Sepal Width	Petal Length	Petal Width	label
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
⋮	⋮	⋮	⋮	⋮	⋮
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

Table 3.1: The iris flower data set containing 150 samples of three related species. All features including sepal length, sepal width, petal length, and petal width are measured in centimetres.

$\hat{e}_1$, $\hat{e}_2$, $\hat{e}_3$, and $\hat{e}_4$ of the sample i as indicated in Table 3.1. Plotting all four features in the same figure is very complicated. However, we can visualise the scatter between any pair of features in the scatterplot and label each species of iris flower in a different colour as shown in Fig. 3.1. From the data set, we may wish to classify the species of iris flowers on the basis of their measured properties. Instead of using all features separately or trying to plot two different features with an ad hoc selection, we can use PCA to reduce the number of dimensions of the data. As a result, the iris flowers can be mapped onto new axes defined by the PCA as $x'_i = \alpha_{1,i}\vec{PC}_1 + \alpha_{2,i}\vec{PC}_2$, where α_1 and α_2 are the eigenvalues of the eigenvectors $\vec{PC}_1$ and $\vec{PC}_2$, which are the first and second principal components containing the highest and second highest variance of the data set. The first two components are defined in terms of the original vectors as

$$\begin{aligned} \vec{PC}_1 &= 0.362\hat{e}_1 - 0.082\hat{e}_2 + 0.857\hat{e}_3 + 0.359\hat{e}_4, \\ \vec{PC}_2 &= 0.656\hat{e}_1 + 0.730\hat{e}_2 - 0.176\hat{e}_3 - 0.075\hat{e}_4. \end{aligned} \tag{3.2}$$

The projection of the original iris data set onto the first two principal components is shown in Fig. 3.2. This plot clearly shows that the transformed data are separable using only the first two components, so the values of two numbers rather than the four numbers stored in the original dataset. Then we can identify the

species of the iris flowers by a classification method, e.g. support vector machine or K-Nearest Neighbours model (which is not covered in this study), which divides up the space plotted in the left panel Fig. 3.2.

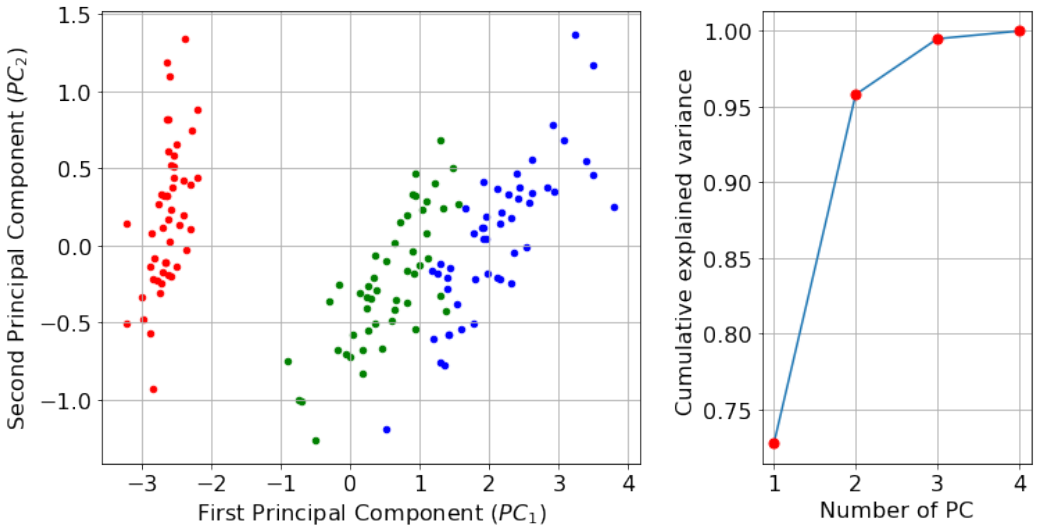


Figure 3.2: Left) The projection of the iris data set onto the first two principal components defined in Equation 3.2; points are colour-coded by their iris family label. Right) The cumulative fractional variance captured by each principal components. The value of variance of first two components combined accounts for 97.8 percent of the total variance.

3.2 The Derivation of Principal Component Analysis

The main propose of this study is to reduce the dimensionality of galaxy spectra, so the notation used in this section will correspond to the structure of the sample set of galaxy spectra. Consider a set of galaxy SEDs, $\{x_i\}$, containing N SEDs with each SED made up of K features (i.e. wavelength bins). We first center the data on the mean of each bin and write the mean-subtracted data as an $N \times K$ matrix,

$$X = \begin{bmatrix} \cdots & x_1 & \cdots \\ & \vdots & \\ \cdots & x_N & \cdots \end{bmatrix}_{N\times K} - \begin{bmatrix} \vdots \\ 1 \\ \vdots \end{bmatrix}_{N\times 1} \begin{bmatrix} \cdots & \bar{x} & \cdots \end{bmatrix}_{1\times K.} \tag{3.3}$$

We can calculate the covariance matrix of the centered data, S_X, by using the formula

$$S_X = \frac{1}{N-1} X^T X, \tag{3.4}$$

where the term $N-1$ is the bias correction arising from the fact the covariances are derived from the data sample. As mentioned in the description of Fig. 3.2, we wish to find a projection of the centered data set that points along the directions of maximal variance. We write the projection of the data as,

$$Y = XW, \tag{3.5}$$

where Y is the matrix of the data projected onto a set of new vectors, W, containing basis vectors, v_i.

$$W = \begin{bmatrix} \vdots & & \vdots \\ v_1 & \cdots & v_N \\ \vdots & & \vdots \end{bmatrix}_{K \times N} \tag{3.6}$$

Each vector v_i in matrix W is chosen to be orthonormal i.e. they satisfy

$$v_i^T v_j = \begin{cases} 1 & ; i = j \\ 0 & ; i \neq j. \end{cases} \tag{3.7}$$

The covariance matrix of the projected data is then

$$\begin{aligned} S_Y &= \frac{1}{N-1} W^T X^T X W \\ &= W^T S_X W. \end{aligned} \tag{3.8}$$

We can find the first principal component (i.e. the unit vector that points along the direction of maximal variance) by maximising the variance using the Lagrangian function. We introduce a new variable, the Lagrange multiplier λ, and add λ times the constraint equation (i.e. Equation 3.7) to the objective equation which is the covariance of the projected data, S_Y, that we want to maximise. The Lagrangian function used to identify the first principal component is written as

$$\mathscr{L}(v_1, \lambda_1) \equiv v_1^T S_X v_1 - \lambda_1 (v_1^T v_1 - 1), \tag{3.9}$$

where the derivatives with respect to λ and W are

$$\frac{\partial \mathscr{L}}{\partial \lambda_1} = v_1^T v_1 - 1, \tag{3.10}$$

$$\frac{\partial \mathscr{L}}{\partial v_1} = 2S_X v_1 - 2\lambda_1 v_1. \tag{3.11}$$

We now can solve the problem above by setting the derivatives to zero and we obtain

$$v_1^T v_1 = 1, \tag{3.12}$$

$$S_X v_1 = \lambda_1 v_1. \tag{3.13}$$

The first principal component, v_1, satisfies Equation 3.12 as it is chosen to be an orthonormal basis vector and Equation 3.13 provides the value of λ_1 as the eigenvalue of the covariance matrix.

$$\lambda_1 = v_1^T S_X v_1 \tag{3.14}$$

The further principal components can be derived in the same way as the first one by using the additional constraint which is the orthogonality between different components (i.e. the case when $i \neq j$ in Equation 3.7). For example, the constraint term for the second component is $\lambda_2(v_2^T v_2)+2\phi v_1^T v_2$. By setting the derivative with respect to v_2 to zero, we will see that ϕ must be zero and we obtain λ_2 which is the second largest eigenvalue associated with the second principal component,

$$\lambda_2 = v_2^T S_X v_2. \tag{3.15}$$

From Equation 3.8, the diagonal values of the covariance matrix S_Y define the amount of variance contained within each principal component (e.g. λ_1 and λ_2 for the first and second components defined in Equation 3.14 and 3.15, respectively). We can define the set of principal components ordered by the variance they are responsible for, with the first component having the most variance.

3.3 Dimensionality Reduction

We have shown that we can find a set of principal components by maximising the amount of variance contained in the components, and that most of the variance is naturally contained in the first few components. It follows that we need only retain the first few components and we can ignore the rest by comparing the cumulative variance in the retained components to the total variance. We can define the fraction of the variance, R, as

$$R \equiv \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{N} \lambda_j} = \frac{\text{sum of the first m variances}}{\text{total variance}}. \tag{3.16}$$

Note that one of the limitations of PCA is that there is no prescription for deciding where to place the cut-off in the retained eigenvectors. This is a subjective choice.

Each galaxy spectrum, x_i, is originally written as a linear combination of vectors corresponding to each wavelength bin and the luminosity in that bin,

$$x_i = \sum_{j=1}^{N} L_{ij} e_j, \tag{3.17}$$

where L_{ij} represents the amplitude (e.g. luminosity) of each $\{e_i\}$ which are {{1, 0, 0, 0, ...}, {0, 1, 0, 0, ...}, ..., {0, 0, ..., 1}}. The original set of galaxy spectra can be written in matrix form as

$$\begin{aligned} x = \begin{bmatrix} \cdots & x_1 & \cdots \\ & \vdots & \\ \cdots & x_N & \cdots \end{bmatrix}_{N\times K} &= \begin{bmatrix} L_{11} & \cdots & L_{1K} \\ & \vdots & \\ L_{N1} & \cdots & L_{NK} \end{bmatrix}_{N\times K} \begin{bmatrix} \vdots & & \vdots \\ e_1 & \cdots & e_K \\ \vdots & & \vdots \end{bmatrix}_{K\times K,} \\ x &= LE. \end{aligned} \tag{3.18}$$

From Equation 3.18, we can readily see that one way to reduce the dimensionality of the original data is by removing columns in matrix E, by applying some form of averaging or smoothing, i.e. by integrating over small features (smoothing and rebinning onto a coarser wavelength grid) or integrating over broad band filters.

However, this approach ignores some features that may have high variance in the data set or either may lose low variance features.

The advantage of PCA is that it provides a better means of dimensionality reduction as the principal components are ordered in terms of how much variance they account for in the data. To reconstruct the original data set x, we use the projection in Equation 3.5 and rename the projection matrix Y as the coefficients or eigenvalues of the principal components α,

$$\alpha = XW. \tag{3.19}$$

Finally we are able to reconstruct the zero-centered data matrix X as

$$[X]_{N\times K} = [\alpha]_{N\times K}[W_{N\times K}]^T \tag{3.20}$$

As the principal components in W are organised by their importance, the data set X can be estimated as

$$[X]_{N\times K} = [\alpha]_{N\times m}[W_{K\times m}]^T, \tag{3.21}$$

where the number of components used, m, is smaller than the original dimension of the data set, K. To make use of Equation 3.21 in our work, we can rewrite each reconstructed spectrum at any given wavelength bin, $x_i(\lambda)$ as a linear combination of the principal components,

$$x_i(\lambda) \approx \bar{x}(\lambda) + \sum_{j=1}^{m} \alpha_{ij} v_j(\lambda), \tag{3.22}$$

where $\bar{x}(\lambda)$ is the mean spectrum, α_{ij} is the coefficient of the principal component, $v_j(\lambda)$ is the principal component vector at the specific wavelength bin λ.

3.4 The Criteria for Choosing the Number of Principal Components

The number of principal components kept determines how well the reconstruction process performs. If too few components are retained then the reconstruction

is inaccurate. On the other hand keeping too many principal components may introduce noise into the reconstruction, which may not worth increasing the number of components. The criteria for selecting the maximum number of principal components to be retained is based on empirical relations derived from different experiments. A particular choice might arise for a given application (see Jolliffe 1986 for a detailed discussion).

The common choice of the number of principal components to retain is made using the variance fraction (i.e. Equation 3.16). An "acceptable" threshold value is set for R such that it captures "most" of the variance in the data set, typically $R = 0.70$ to 0.95. In some cases, a substantial number of principal components may be required to reach the threshold. The change in the gradient of the variance with number of retained eigenvectors (i.e. the knee in the scree plot, see left plot of Figure 3.2) could be used instead (Cattell 1966). However, in reconstructing galaxy SEDs, we place requirements on the accuracy of the reconstructed spectra over the whole wavelength range rather than the amount variance captured by principal components. The criteria used in this work will be discussed separately in the next chapter.

3.5 Application of PCA to Spectra

The PCA technique has been used widely to study spectral classification and to infer physical properties using galaxy spectra (e.g., Connolly et al. 1995; Folkes et al. 1996; Folkes et al. 1996). Madgwick et al. (2003) found that the linear combination of the first two principal components resulting from an analysis of spectra in the 2dF Galaxy Redshift Survey (2dFGRS) is correlated with morphological type and has a tight correlation with the star formation rate (e.g. Madgwick et al. 2003 and also Ronen et al. 1999 for a similar result). Chen et al. (2012) used PCA to estimate the physical properties of galaxies from the Baryon Oscillation Spectroscopic Survey (BOSS) and found that 7 principal components provide a good fit to

the original spectra. Moreover, the prediction of synthetic galaxy spectra from a population synthesis model can be speeded up by training a neural network (NN) with the decomposed data from the PCA. Alsing et al. (2020) found that, instead of training the NN to reproduce the whole spectrum with several thousand spectral features (i.e. wavelength bins), training it to produce only a few tens components provides a great improvement in accuracy while the calculation is much faster than calculating the spectra with direct SPS modeling once the NN is trained.

Chapter 4

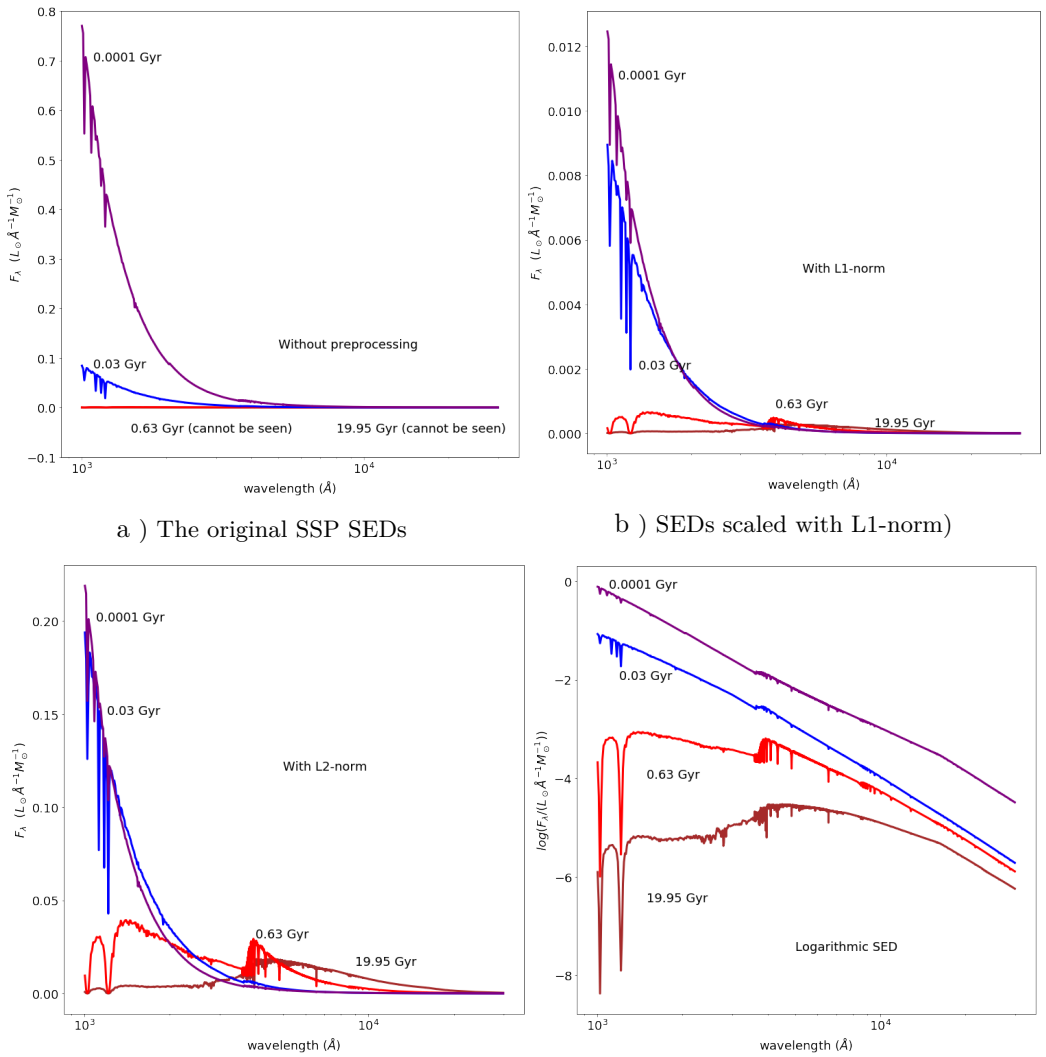
Results I: The PCA of Simple Stellar Populations

We introduced a tool for data compression in the previous chapter and have shown that the PCA technique can be helpful in studying galaxy SEDs. In this chapter we will apply PCA to the SEDs of simple stellar populations i.e. those defined by a fixed age and metallicity. First, we will describe the method of preprocessing of the simple stellar population SEDs before being the PCA is applied. This additional step is necessary because of the large range of values covered by the spectra. Then we will show the results of the reconstruction of the SSP spectra.

4.1 Data Preparation of the Simple Stellar Population Spectra for PCA

From Fig. 2.2a, we can see that the SEDs of simple stellar populations change dramatically at very young ages, and start to change more gradually when they become very old. The PCA technique is very sensitive to outliers in the data set. One significant outlier may lead to a poor overall result since the PCA tends to fit the outlier well (Serneels and Verdonck 2008). In our SSP SED sample, the dynamic range of the fluxes seen in very young populations can be viewed as

outliers as we can see in Fig. 4.1a; they are very different from other SSP outputs. The flux changes from the order of 10^0 at short wavelengths for the youngest SSP age to the order of 10^{-5} for the oldest age output.



a) The original SSP SEDs

b) SEDs scaled with L1-norm)

c) SEDs scaled with L2-norm

d) Logarithmic SEDs

Figure 4.1: A visualisation of the SSP SEDs. The value on the y-axis is plotted on a linear scale. a) The original SSP SEDs clearly contain a huge dynamic range, which are the specific flux of young populations at wavelength $\sim$ 1000Å, in the sense that they can be seen as distinctly separate from the other SEDs on a linear scale. Compare this with the same information shown on a logarithmic scale in Fig. 2.2a. b) - c) The SSP SEDs again plotted on a linear scale, but scaled by using the L1 and L2 normalisation, respectively (see text). d) The SSP SEDs plotted on a logarithmic scale. The outliers evident in panel a) are now less extreme in panel b) and c), especially in logarithmic scale.

4.1.1 The L^P-norm Normalisation of Spectra

To reduce the severity of the outliers (or the dynamic range) in the data set for the PCA, we rescale each individual SED by dividing it by the L^P-norm value. The norm is defined as

$$\begin{aligned}\eta_i =& ||f_{i,\lambda}||_P = \left(\sum_{\lambda_j}^{\lambda_K} f_{i,\lambda_j}^P\right)^{1/P} \\ =& (f_{i,\lambda_1}^P + f_{i,\lambda_2}^P + ... + f_{i,\lambda_K}^P)^{1/P},\end{aligned} \tag{4.1}$$

where f_{i,λ_j} is the flux in wavelength bin λ_j of spectrum i. As a result of applying this normalisation, the scaled SEDs shown in the panel b) and c) of Fig. 4.1c showing less extreme outliers. The change in flux seen for the youngest SSPs is now ~ 1 dex instead of 5. Hence we now can redefine the data matrix from Equation 3.3 to be the normalised data matrix, X', by multiplying the inverse of the diagonal matrix containing the L^P-norm values of each spectrum as

$$X' = \begin{bmatrix} \cdots & x'_1 & \cdots \\ & \vdots & \\ \cdots & x'_N & \cdots \end{bmatrix} - \begin{bmatrix} \vdots \\ 1 \\ \vdots \end{bmatrix} \begin{bmatrix} \cdots & \bar{x}' & \cdots \end{bmatrix}, \tag{4.2}$$

$$X' = \mathrm{diag}\{\eta_1, \eta_2, ..., \eta_N\}^{-1} \begin{bmatrix} \cdots & x_1 & \cdots \\ & \vdots & \\ \cdots & x_N & \cdots \end{bmatrix} - \begin{bmatrix} \vdots \\ 1 \\ \vdots \end{bmatrix} \begin{bmatrix} \cdots & \bar{x}' & \cdots \end{bmatrix} \tag{4.3}$$

where $\bar{x}'$ is the mean of the normalised spectra. The normalised data matrix X' then replaces the original data matrix in Equation 3.4 and the PCA will define the new eigenvectors (principal components) based on this normalised data.

By substituting the normalised matrix data X' into Equation 3.4, a spectrum for the simple stellar population can be approximately described by

$$x_i(\lambda) \approx \eta_i \left[\bar{x}'(\lambda) + \sum_{j=1}^{m} \alpha_{ij} v'_j(\lambda)\right]. \tag{4.4}$$

4.1.2 The Logarithm of Spectra

The change seen in the spectra at young ages can be reduced by rescaling using the L2-norm normalisation technique. However, the change in flux is still on the order of one magnitude in the new units. We can clearly see this if we plot the spectra on a logarithmic scale. The change is now of the same order of magnitude (e.g. compare Fig. 4.1c with Fig. 4.1d.) Therefore, we can also apply the PCA to the logarithm of the spectra. Equation 3.22 becomes

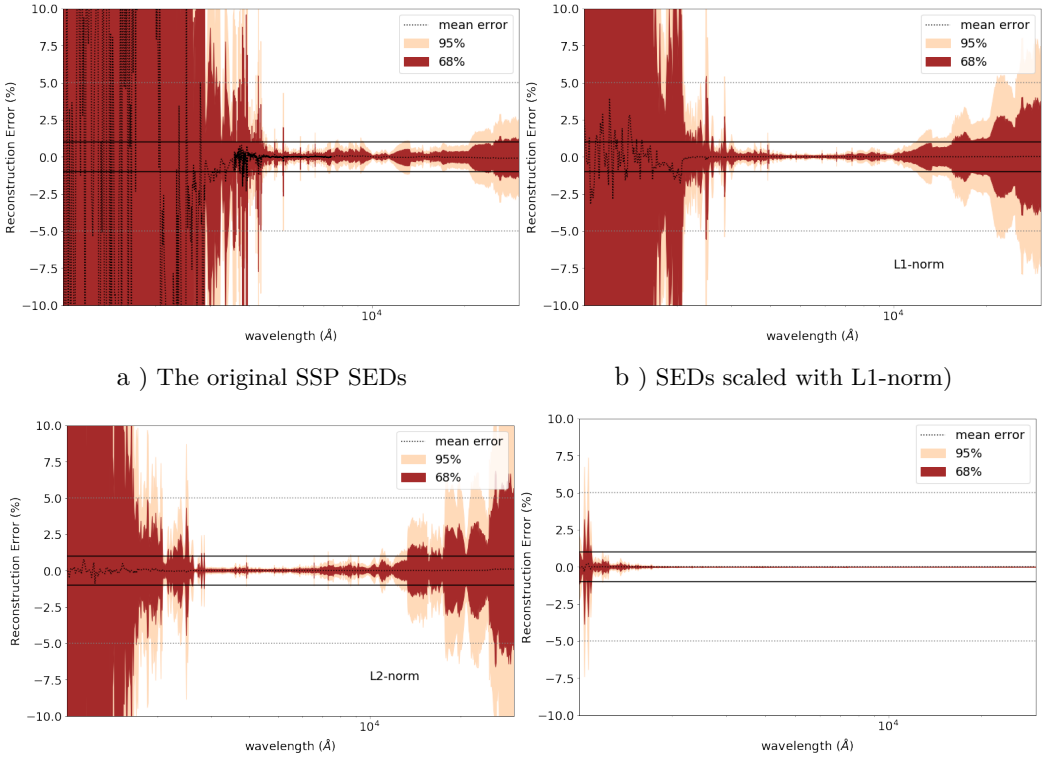
$$\log x_i(\lambda) \approx \bar{x}'_{log}(\lambda) + \sum_{j=1}^{m} \alpha_{ij} v'_{log,j}(\lambda), \tag{4.5}$$

where $\bar{x}'_{log}(\lambda)$ is the mean of the logarithmic spectra and $v'_{log,j}$ are the principal components of the logarithmic spectra.

4.1.3 Comparison Between Different Normalisation Techniques

The dynamic range in the data set can be reduced by rescaling the SEDs using the L^P-norm normalisation or by taking the logarithm of the spectra. By doing so, the outliers become less extreme in the new units which prevents the PCA from being unduly affected by them. To demonstrate this, we can compare the results of the PCA applied to the data set with different normalisation methods, as shown in Fig. 4.2 (the details behind the creation of this plot will be discussed in Section 4.2).

Fig 4.2 shows a comparison between the results of the PCA when applied to different preprocessing techniques for solar metallicity SSP SEDs at different ages. Even though the PCA performs very well when applied to the logarithmic spectra, it is worth mentioning that the objective of this study is to calculate composite stellar populations via the stellar population synthesis approach, which relies on the linear superposition of the SSPs. And we wish to reduce the dimensionality of the SEDs by mapping the SEDs onto the principal components obtained from the PCA. The normalisation method used in the data preprocessing step therefore



a) The original SSP SEDs

b) SEDs scaled with L1-norm)

c) SEDs scaled with L2-norm

d) Logarithmic SEDs

Figure 4.2: A comparison of the SSP SEDs reconstruction results by applying the PCA with different normalisation techniques (the original spectra, the L1-norm normalised spectra, the L2-norm normalised spectra, and the logarithm of the spectra). The plots show the reconstruction errors of SSP SEDs when using 50 principal components as a function of wavelength. The black dotted lines represent the mean of the reconstruction error. The regions show the 68% (brown) and 95% (cream) percentiles of the deviation from the mean. The 1% and 5% error intervals are shown by the horizontal solid and dotted lines.

needs to be linear (i.e. the values of specific flux at all wavelength bins of a spectrum can only be multiplied by a scalar). Non-linear normalisation methods (e.g. taking the logarithm of the spectrum) will lead to a complexity in calculating the SED that will negate any reduction in dimensionality. Therefore, in this study we will henceforth apply the PCA to the L2-norm scaled spectra as it provides the "best" reliability of the SSP SED calculation.

4.2 Principal Component Analysis of a Fixed Metallicity Simple Stellar Population

In Chapter 2, we introduced the stellar population synthesis model and discussed the data preprocessing technique in Section 4.1. In the application of the PCA to the SEDs of the simple stellar populations, we aim to reproduce the SSP spectra over the whole wavelength range, from the UV to the near IR. Firstly, we will consider a less complex data set which is the SEDs for a fixed solar metallicity SSP.

4.2.1 The Solar Metallicity SSP: PCA applied to the whole wavelength range at once

Originally the SSPs computed from the FSPS code (Foreman-Mackey et al., 2014) with the inputs as shown in Table 2.1 have 107 ages that cover the 10^{-4} to $10^{1.3}$ Gyr in logarithmic steps. This parameter space will be our goal for this first phase of the SSP SED reconstruction. By applying the PCA to this data set, the mean spectrum and the principal components of the solar metallicity SSPs with the original age grid are shown in Fig. 4.3.

The first eigenspectrum (first principal component: PC_1) captures most of the variance in the data set. The subsequent eigenspectra are then identified to have the second highest amount of variance captured and so on. The values of explained variances (obtained from Equation 3.12), are plotted in Fig. 4.4, which shows the scree plot of the variance and the fractional variance contained in the first 10 components.

A typical choice for selecting the number of principal components to keep is made by setting the threshold of the cumulative explain variance ratio at $\sim 70\% - 95\%$ as we have discussed in Section 3.4. From Table 4.1, the first 3 components together combined capture 97.78% of the total variance. However, the

PC	EV	%EVR	C. %EVR
1	0.337	83.33	83.33
2	4.67E-2	11.57	94.90
3	1.16E-2	2.876	97.78
4	5.79E-3	1.431	99.21
5	1.60E-3	0.3966	99.61
6	6.3eE-4	0.1566	99.76
7	4.64E-4	0.1149	99.88
8	1.65E-4	0.0408	99.92
9	1.25E-4	0.0310	99.95
10	6.20E-5	0.0154	99.96

Table 4.1: Values of explained variance (EV; from Eqn. 3.14), explained variance ratio percentage (%EVR*), and the cumulative explained variance ratio (C. %EVR; from Eqn. 3.16) of the first 10 principal components. *Note that the value of total explained variance is 0.4043.

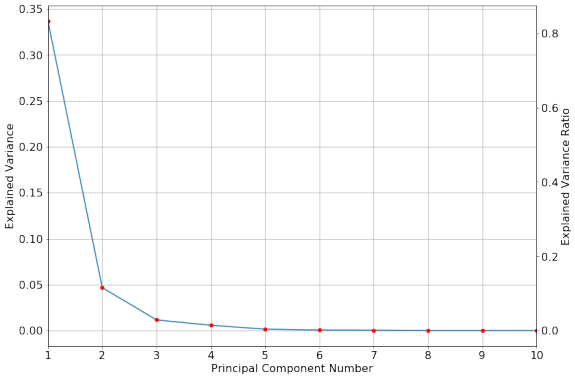


Figure 4.4: The explained variances and the explained variance ratios captured by the first 10 principal components of the fixed-solar metallicity SSP SEDs

reconstructed SSP SEDs with 3 components are not sufficient to represent the original SEDs. With such a low number of principal components we are only able to provide a good reproduction of the shape of the SSPs at ages around 10 Myr (Fig. 4.5b) but the reconstruction error across all wavelength bins is relatively high. Moreover, with this number of principal components, the reconstruction completely fails to rebuild very young SSPs in the infrared region (Fig. 4.5a) and fails to reconstruct the ultraviolet spectra of old age SSPs (Fig. 4.5c and Fig. 4.5d).

In our case, the number of components to be retained is driven by the accuracy of the reconstructed SEDs. If we wish to rebuild the SSP SEDs with a typical accuracy level such that all specific flux lies within an error of $< 5\%$, 77 principal components are required. Moreover, for the case of "extremely" accurate reconstruction ($< 1\%$ error), we need 91 components. See Fig. 4.6 for the distribution of the reconstruction error as a function of wavelength, which is the criteria for how many principal components we need to use.

As shown in Fig. 4.6, The PCA applied over the whole wavelength range can

reconstruct the SEDs in the optical very well compared to the reconstruction in the UV and NIR. Because of this we can decide to apply the PCA separately to different wavelength ranges in an attempt to improve the accuracy in the UV and IR, at the expense of a modest reduction in the accuracy in the optical.

4.2.2 The Solar Metallicity SSP: PCA applied to distinct wavelength ranges

PCA applied in one go to the full wavelength range tends to result in better reproductions of the spectra in the optical rather than in the UV and NIR. We now apply the PCA by dividing the SSP spectra into three wavelength ranges. The wavelength ranges are the UV (1000 - 3500 Å), the optical (3500 - 7500 Å), and the NIR (7500 - 30000 Å). The sample size of the SSP SEDs is the same as used for the whole wavelength range PCA. As a result, we are able to reduce the number of principal components needed to reconstruct the SSP SEDs at both accuracy levels quoted above in every wavelength bin. Fig. 4.7 shows a comparison between the reconstruction error distribution for the PCA applied to the whole wavelength range and the PCA applied to the SEDs divided into three wavelength ranges. The boundaries between each band are plotted at 3500Å and 7500 Å with blue and orange short vertical lines at the bottom of the plots. By dividing the wavelength space into these three ranges, we only need 68 principal components in total (45 for the UV, 14 for the optical, and 9 for the IR) to reconstruct the SEDs in the typical accuracy case ($< 5\%$ error). 53 components for the UV, 20 for the optical, and 12 for the NIR (85 PCs in total) are needed to reconstruct the SEDs in the extremely accurate case ($< 1\%$ error). In comparison to the whole wavelength range PCA, with this procedure we can reduce the total number of components from 77 to 68 for the typical case and from 91 to 85 for the extreme case.

To show the solar metallicity SSP reconstructed using the results of the PCA, we provide a set of examples at four different ages that show a considerable change in the spectrum: 0.2 Myr, 10 Myr, 1.0 Gyr, and 19.95 Gyr (see Fig. 4.8 and 4.9).

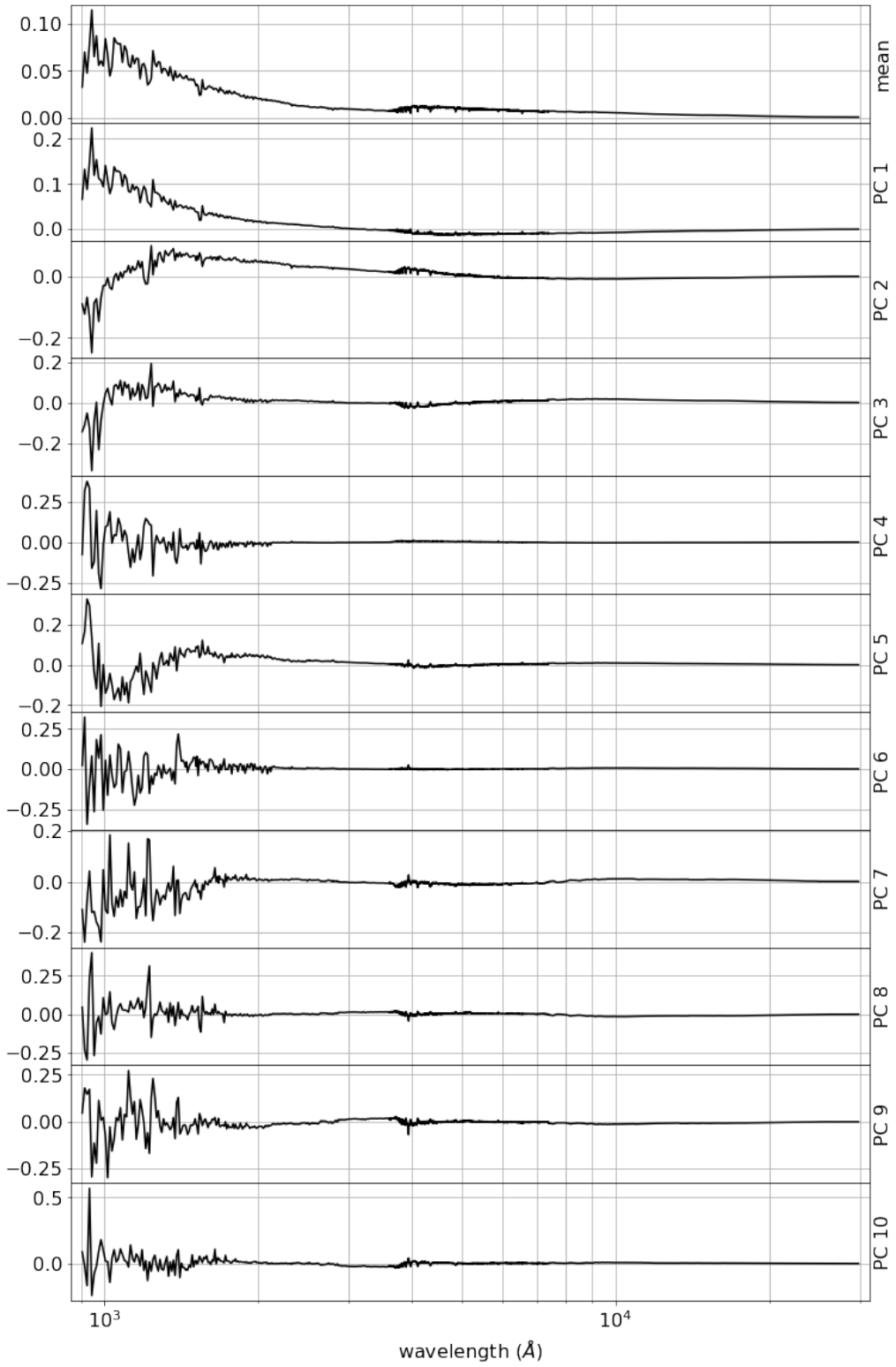
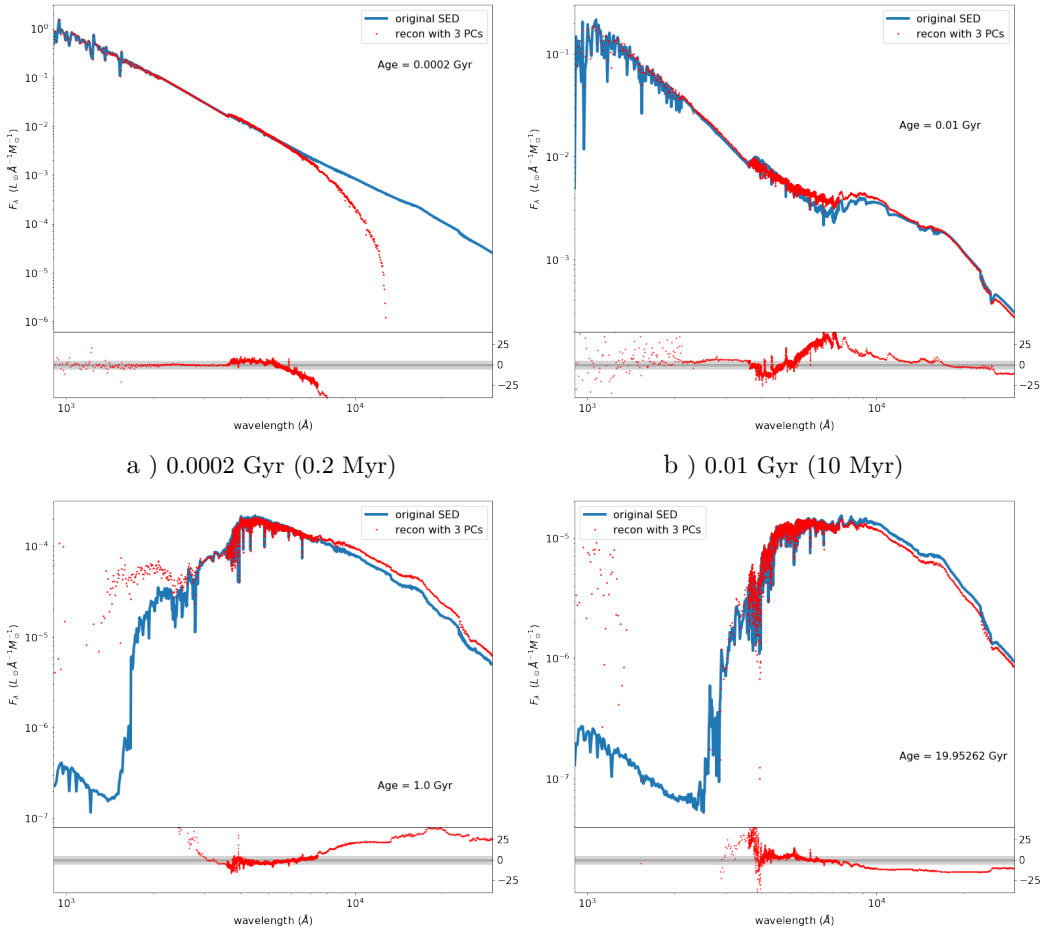


Figure 4.3: The mean of the normalised spectra (top panel) and the first five principal components of the solar-metallicity SSP SEDs with the original age grids. The components are listed in increasing order of component number from top to bottom.



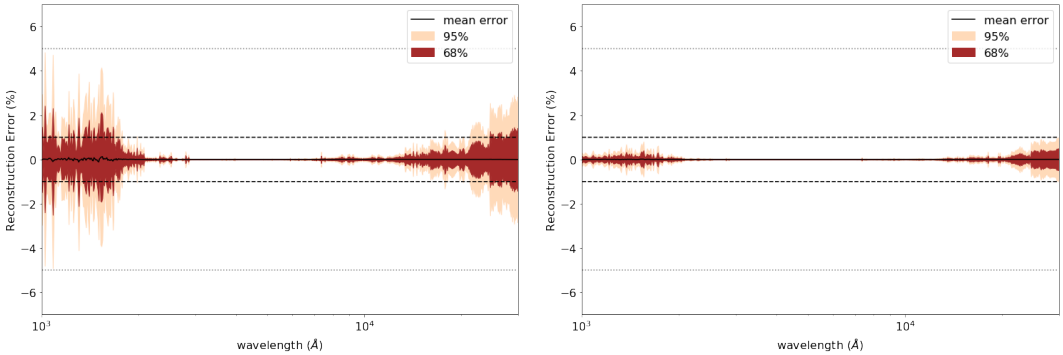
a) 0.0002 Gyr (0.2 Myr)

b) 0.01 Gyr (10 Myr)

c) 1.0 Gyr

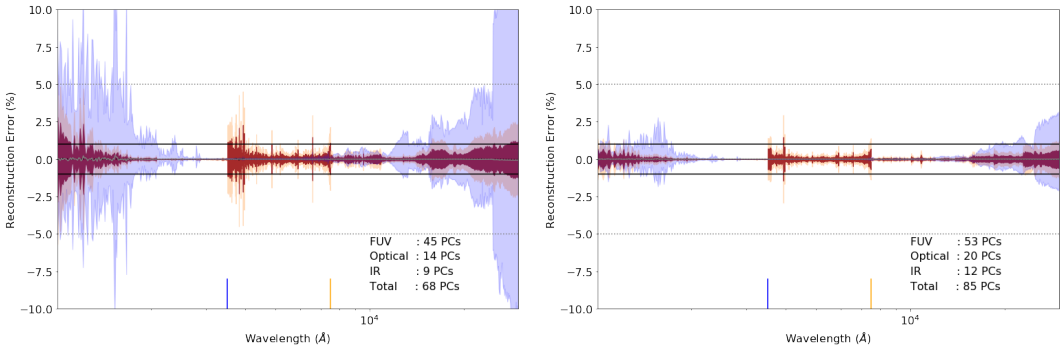
d) 19.95 Gyr

Figure 4.5: The SED reconstruction of the solar metallicity SSPs using 3 principal components at 4 different ages (0.2 Myr, 10 Myr, 1.0 Gyr, and 19.95 Gyr). The blue solid lines represent the original SEDs obtained from the SPS model and red dots show the reconstructed SSPs using 3 components. The percentage of the reconstruction error is shown as red dots in the inset panel at the bottom of each plot.



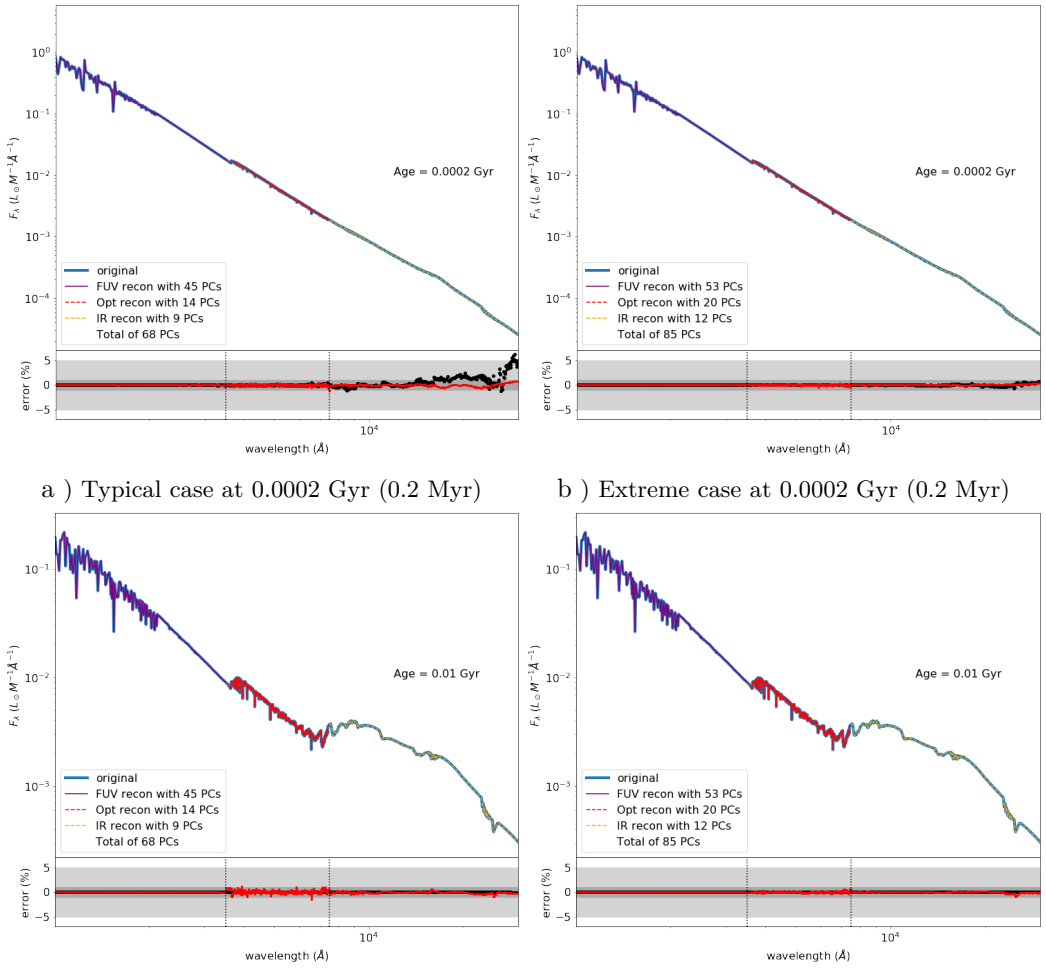
a) Typical Accuracy ($< 5\%$) with 77 PCs b) Extreme Accuracy (< 1 %) with 91 PCs

Figure 4.6: The distribution of reconstruction error as a function of wavelength. The black solid lines are the mean of the reconstruction error. Brown and cream shading represent the 1 and 2 standard deviation from the mean. The 1% and 5% reference errors are shown by dashed- and dotted-lines. a) 77 components are required to rebuild the SSP SEDs to within 5% error over the whole wavelength range for all ages. b) In a case of 1% error, 91 components are required.



a) Typical error case ($< 5\%$) with 68 PCs b) Extreme error case (< 1 %) with 85 PCs

Figure 4.7: The distribution of reconstruction error as a function of wavelength computed separately in 3 different bands. The same description as in Fig. 4.6 plus light blue color referring to the 2 standard deviation from the mean of the whole spectrum PCA. See text for description.



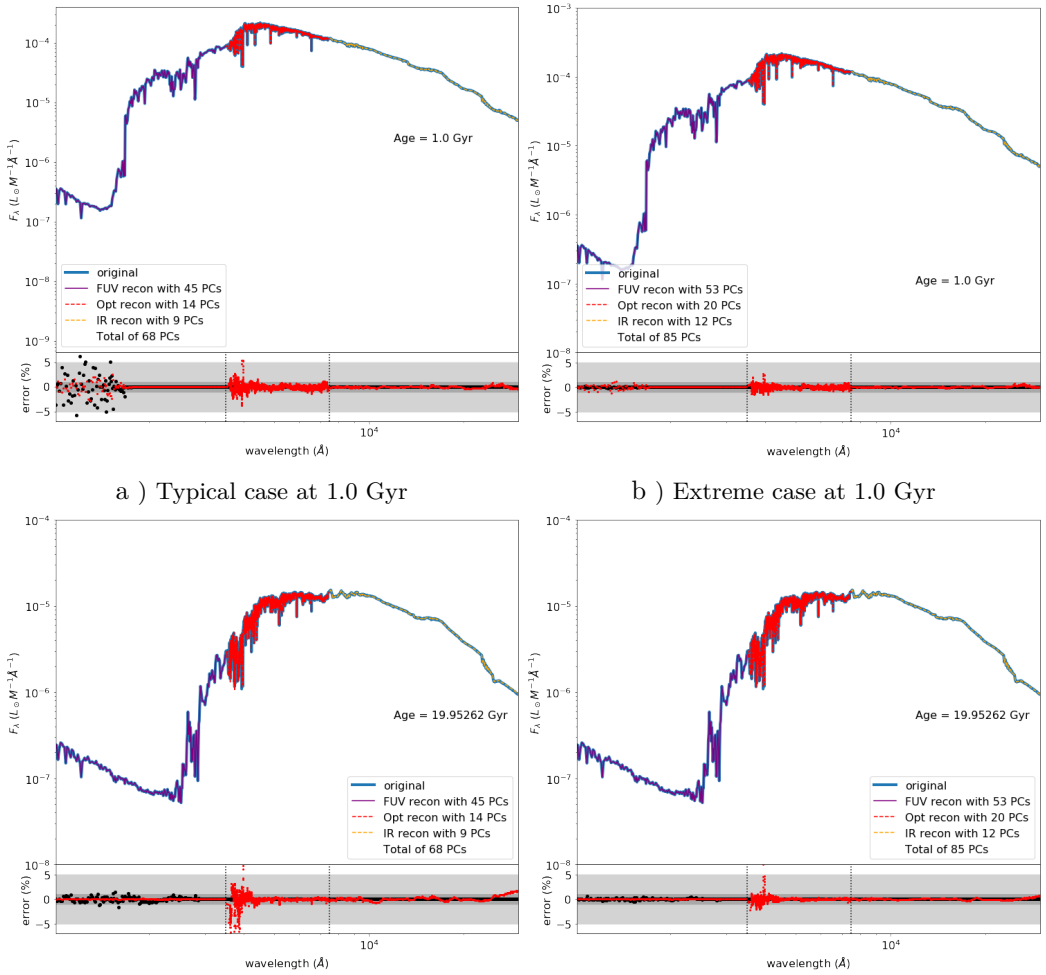
a) Typical case at 0.0002 Gyr (0.2 Myr)

b) Extreme case at 0.0002 Gyr (0.2 Myr)

c) Typical case at 0.01 Gyr (10 Myr)

d) Extreme case at 0.01 Gyr (10 Myr)

Figure 4.8: The SED reconstructions at the age of 0.2 and 10 Myr by using the PCA of UV, Optical, and NIR bands. The solid blue line represents the original SED. Purple, red, and orange lines are for UV, optical, and NIR. In the bottom panel, the red dots show the percentage error of the separate wavelength PCA whilst black dots show that of the whole-wavelength-range PCA by using the same total number of components.



a) Typical case at 1.0 Gyr

b) Extreme case at 1.0 Gyr

c) Typical case at 19.95 Gyr

d) Extreme case at 19.95 Gyr

Figure 4.9: The SED reconstructions at ages of 1.0 and 19.95 Gyr by using the PCA of UV, Optical, and NIR bands. The solid blue line represents the original SEDs. Purple, red, and orange lines are for UV, optical, and NIR. In the bottom panel, the red dots show the percentage error of the separate wavelength PCA whilst black dots show that of the whole wavelength range PCA by using the same total number of components.

4.3 Principal components of the SEDs of the Simple Stellar Populations With Varying Age and Metallicity

In the previous section we showed the result of the PCA for SSPs with fixed solar metallicity and varying ages. In this section we apply the PCA to a more complicated data set. Instead of applying the PCA to a fixed metallicity SSP spectra, we now vary both age and metallicity.

4.3.1 Sample Size of the Simple Stellar Population SEDs

In the application of PCA to the SEDs of simple stellar populations, we aim to reproduce the SSP spectra of the whole parameter space. According to the original parameter space provided by the FSPS code, the time grid covers the range from 10^{-4} to $10^{1.3}$ Gyr with logarithmic spacing and the metallicity grid covers $\log(Z/Z_{\odot}) = -2.5$ to 0.5. In this work we will keep the original parameter space without adding another ages and metallicities. The summary of parameter space is shown in Table 4.2.

Parameters	Coverage
Age of SSPs	logarithmically distributed from 10^{-4} to $10^{1.3}$ Gyr with 107 bins
Metallicity	logarithmically distributed from $-2.5 \leq \log(Z/Z_{\odot}) \leq 0.5$ with 12 values
Wavelength bins	In rest-frame from 1000Å to 30,000Å ($\sim$ FUV to K band)

Table 4.2: The summary of parameter grids for calculating the SSP SEDs. The total number of SEDs is 1284 spectra.

4.3.2 Simple Stellar Population SEDs Reconstruction

We use the same technique as used in calculating the principal components for the fixed metallicity SSPs, namely to apply the PCA to the three different wavelength ranges separately (See Section 4.2.2). The first 10 principal components of the UV, optical, and IR PCA are shown in Fig. 4.10, 4.11, and 4.12.

Once the principal components are defined, the spectrum of an SSP can be reconstructed by using Equation 4.4. As we also discussed about the number of components to keep in Section 4.2.1, we experience a similar situation in the case of the SSPs with varying metallicities and ages. Specifically, the first 2 components of the whole wavelength range PCA of this data set capture $\sim 98\%$ of the total variance of the spectra but they cannot be used by themselves to represent the SSP SED to an acceptable level of accuracy. The demonstration of the reconstruction error can be represented as a distribution of the error of the whole parameter space as a function of wavelength. Fig. 4.15 shows the $2-\sigma$ range of the error distribution as a function of wavelength with the same description in Fig. 4.7. In this case, 26 components of the optical PCA and 12 components of NIR PCA are selected to rebuild the spectra to reach 5% accuracy compared with the 50 and 24 that are needed to reach 1% level in the wavelength ranges. However, the number of components of the UV range goes up to more than 100 components yet still does not narrow the $2-\sigma$ range of the error distribution of the whole parameter space to be less than 5%. We will keep the number of the principal components used in the UV as a free parameter in calculating the composite stellar population SED in the next Chapter.

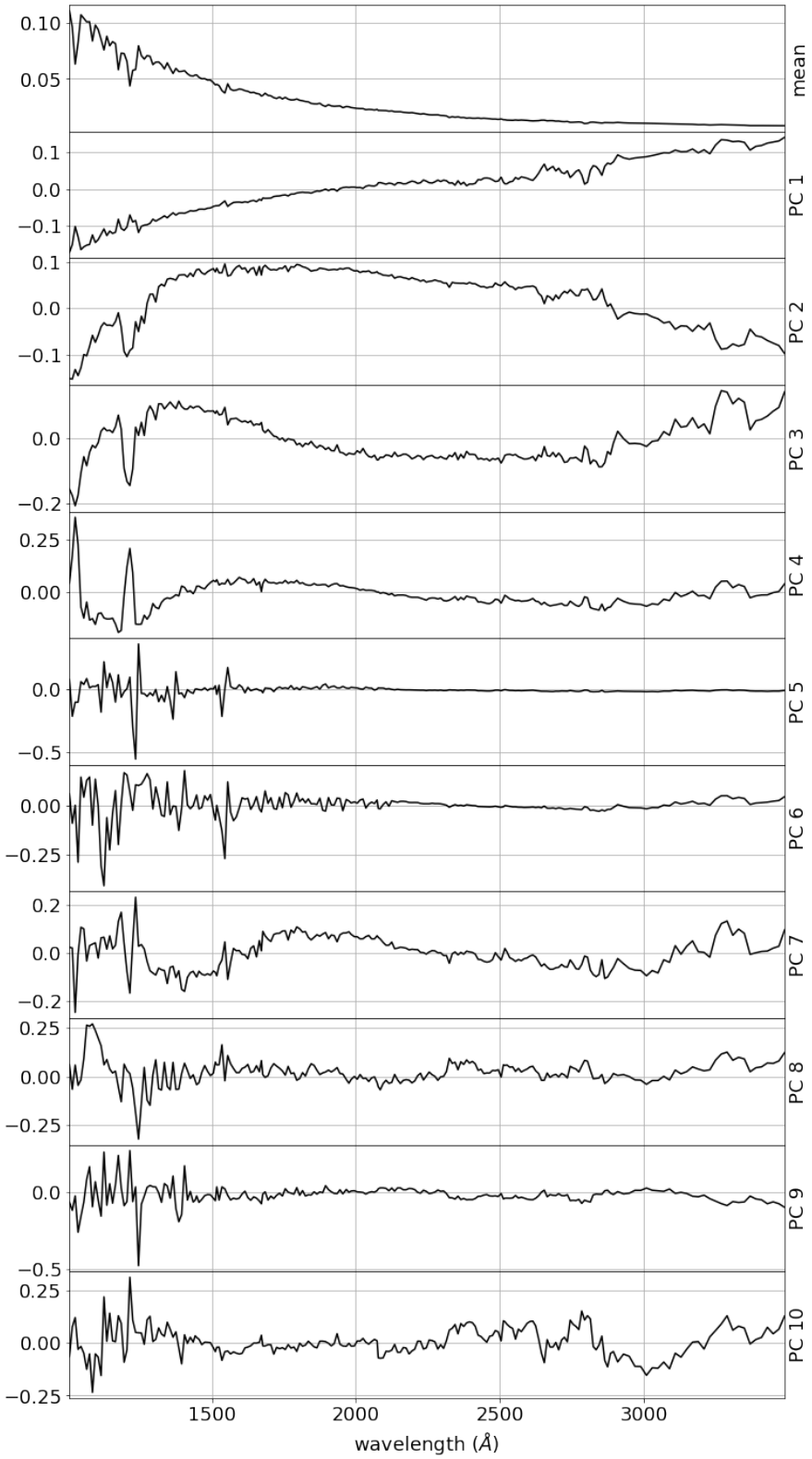


Figure 4.10: The mean of the normalised UV spectra (top panel) and the first 10 principal components of the SSP SEDs (listed from top to bottom). The wavelength range covers 1000 to 3500 Å.

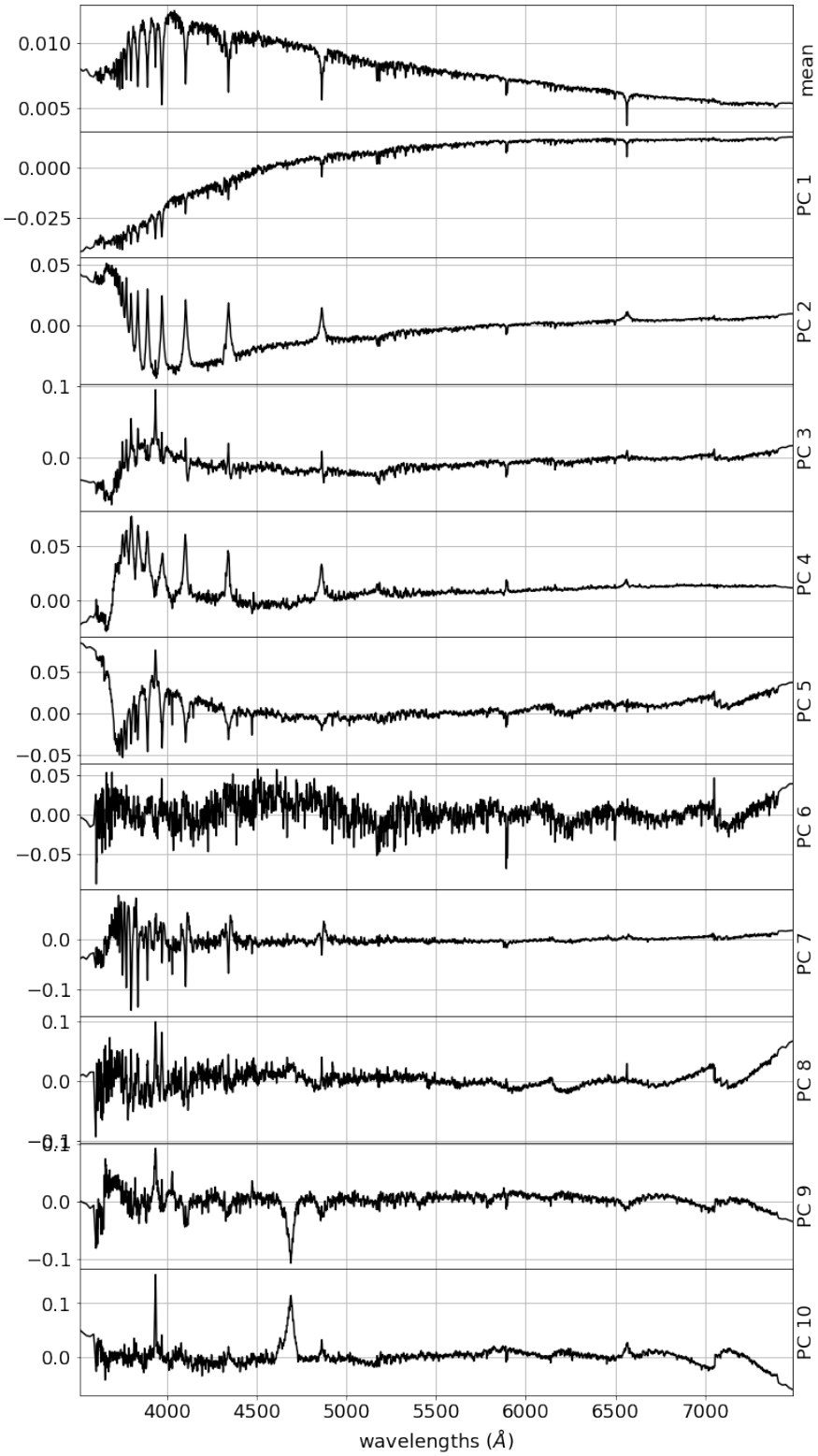


Figure 4.11: The mean of the normalised optical spectra (top panel) and the first 10 principal components of the SSP SEDs (listed from top to bottom). The wavelength range covers 3500 to 7500 Å.

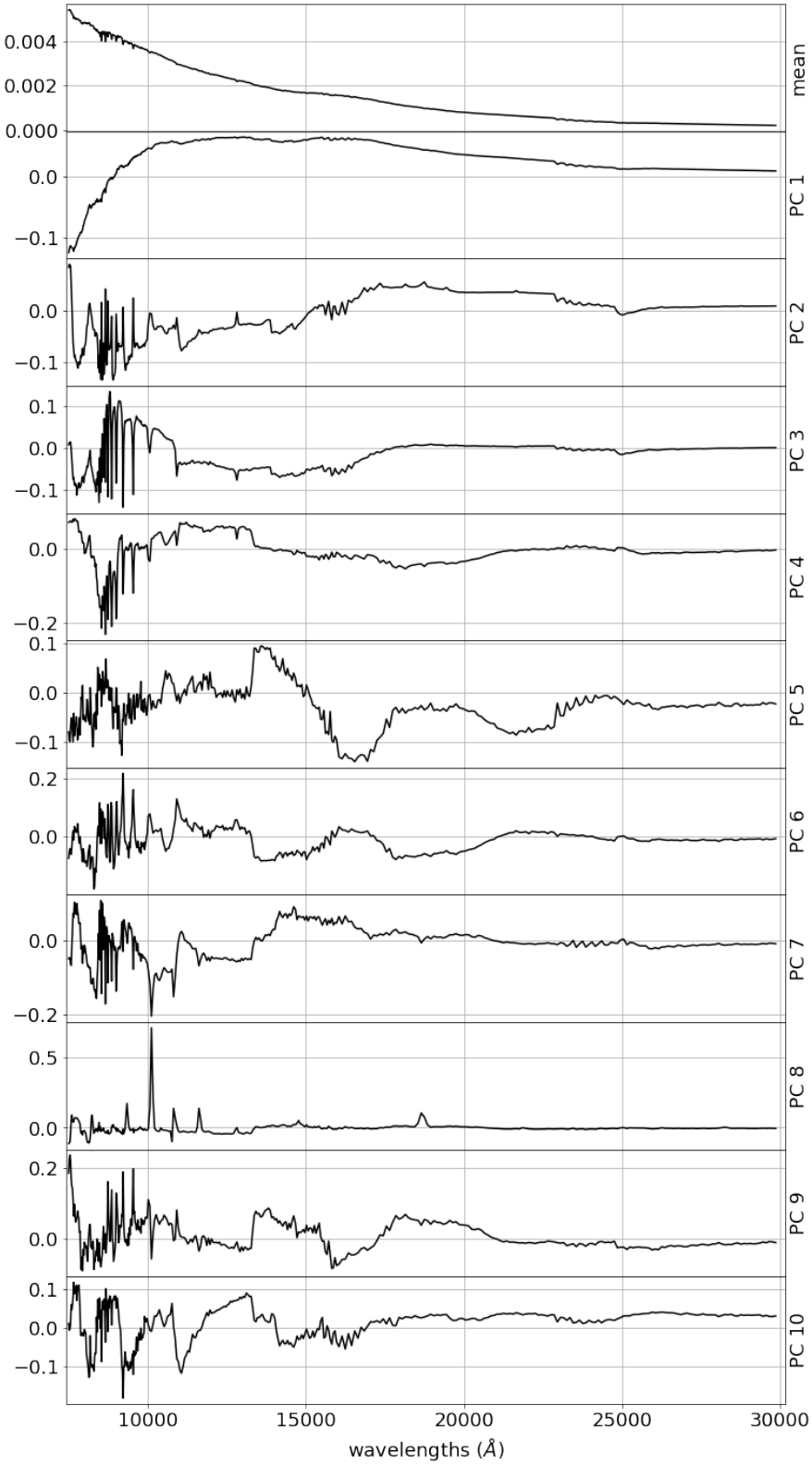
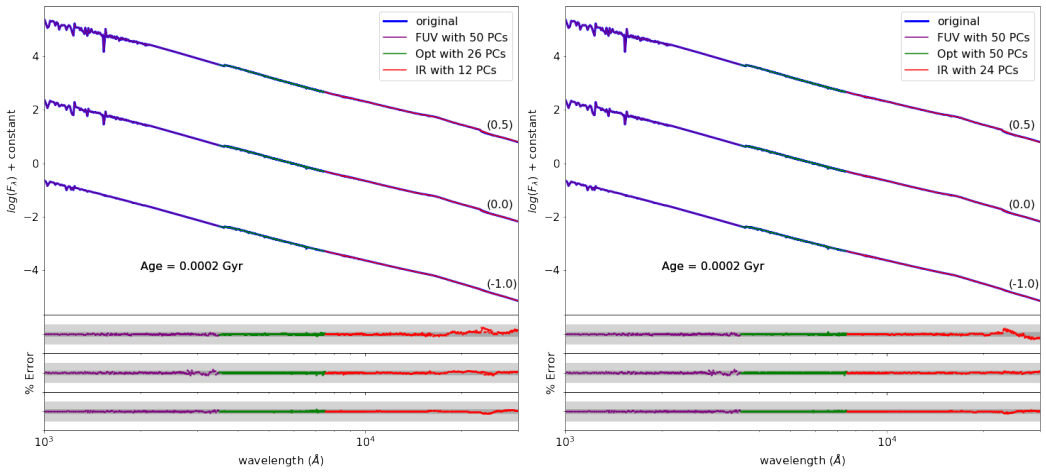
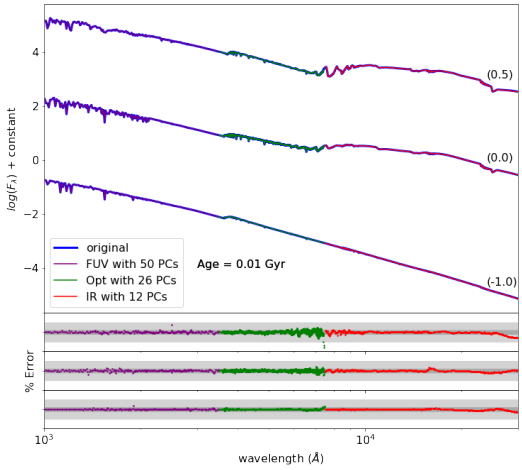


Figure 4.12: The mean of the normalised NIR spectra (top panel) and the first 10 principal components of the SSP SEDs (listed from top to bottom). The wavelength range covers 7500 to 30000 Å.

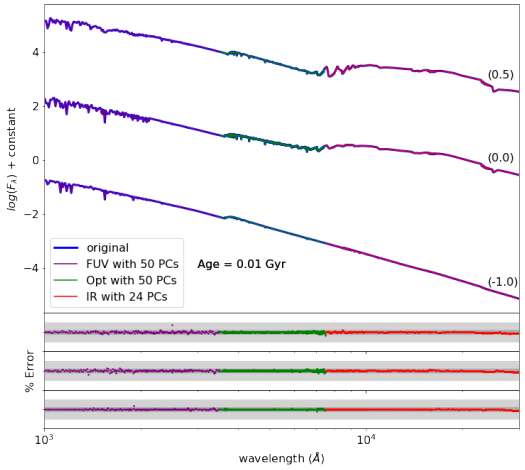


a) Typical case at 0.0002 Gyr (0.2 Myr)

b) Extreme case at 0.0002 Gyr (0.2 Myr)

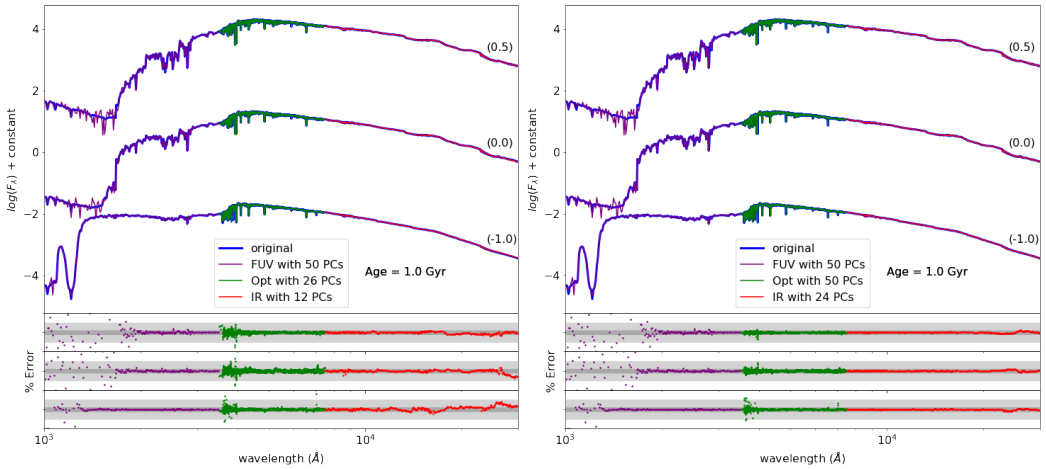


c) Typical case at 0.01 Gyr (10 Myr)



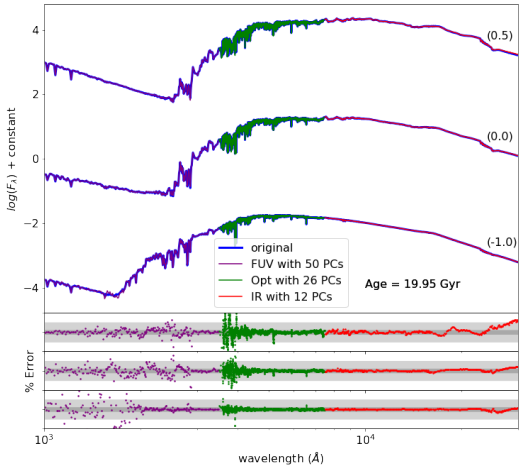
d) Extreme case at 0.01 Gyr (10 Myr)

Figure 4.13: Same description as in Fig. 4.8 but for the SSPs at 10 Myr, 1.0 Gyr, and 19.95 Gyr. For each panel, the values of $\log(Z/Z_{\odot})$ are 0.5, 0.0, and -1.0 from top to bottom as shown in the brackets.

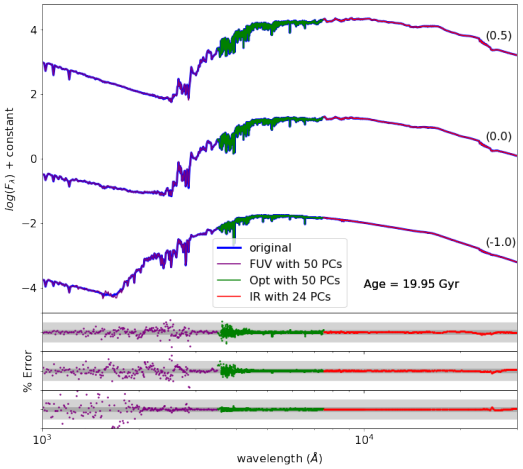


a) Typical case at 1.0 Gyr

b) Extreme case at 1.0 Gyr

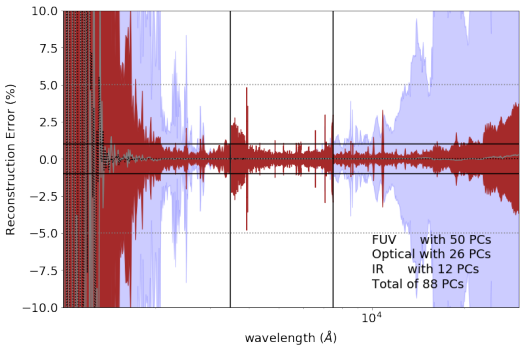


c) Typical case at 19.95 Gyr

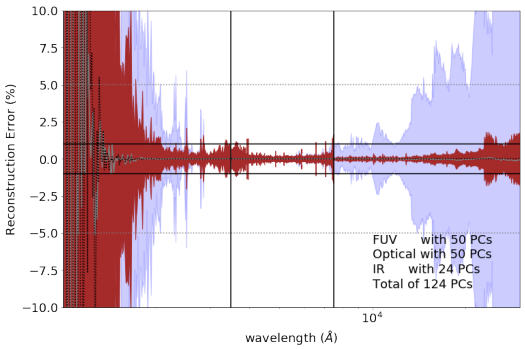


d) Extreme case at 19.95 Gyr

Figure 4.14: Same description as in Fig. 4.13 but for the SSPs at 10 Myr, 1.0 Gyr, and 19.95 Gyr.



a) < 5% error with total of 88* PCs



b) < 1 % error with total of 124* PCs

Figure 4.15: The distribution of reconstruction error as a function of wavelength computed separately in 3 different ranges. The same ranges as shown in Fig. 4.15. *See text for the explanation of the total number of components.

Chapter 5

Results II: The Composite Stellar Population from PCA

5.1 Calculating the composite stellar population using PCA

We introduced stellar population synthesis models in Chapter 2 and computed the principal components of the simple stellar populations which are the building blocks of the composite stellar population in Chapter 4. In this chapter we will compute the spectra of composite stellar populations by replacing the simple stellar populations with the result of the PCA from Chapter 4.

We have mentioned the method used for the CSP calculation that differs from the direct FSPS calculation in Equation 2.6. And the SED of a simple stellar population that has been decomposed using PCA has the form

$$\mathcal{L}_{SSP,i} = \eta_i \left[\mu + \sum_{j=1}^{m} \alpha_{i,j} v_j \right], \tag{5.1}$$

where η_i is a normalisation factor, μ is the mean spectrum (of the SSPs), and $\alpha_{i,j}$ is the coefficient of the principal component v_j. The value of m is the number of principal components used to represent the SSP. The reason we need to multiply

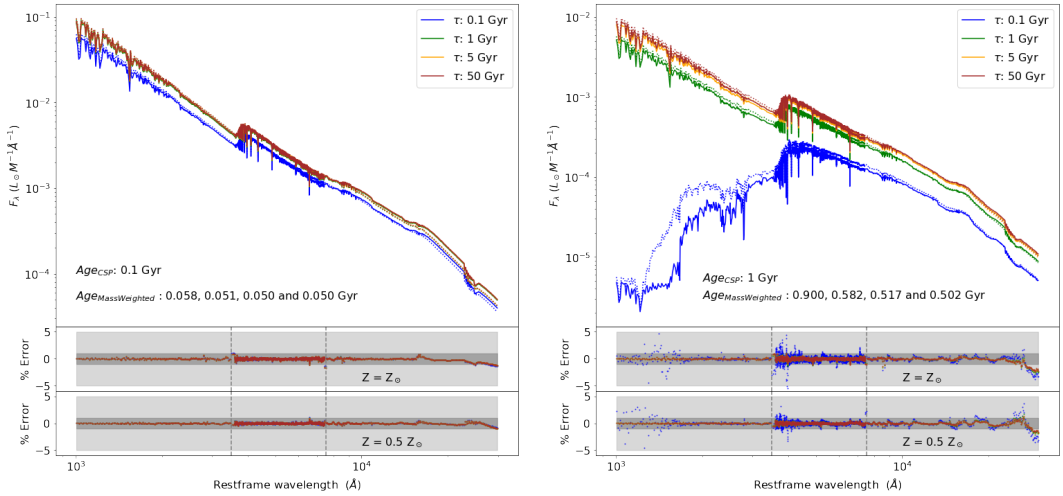
the reconstructed SED by the normalisation factor η_i is that we are applying the PCA to the re-normalised spectra. In the CSP calculation we need to convert the reconstructed spectra back to the original units. Then, by substituting this back into Eq. 2.6, we can compute the CSP spectrum by using a linear combination of the principal components of all SSPs as

$$\begin{aligned}\mathcal{L}_{CSP}(t_{age}) &= \sum_{i=1}^{n_{tage}} \mathcal{L}_{SSP,i}\ w_i \\ &= \sum_{i=1}^{n_{tage}} \left\{ \eta_i \left[\mu + \sum_{j=1}^{m} \alpha_{i,j} v_j \right] \right\} w_i \\ &= \mu \sum_{i=1}^{n_{tage}} \eta_i w_i + \sum_{j=1}^{m} \left[\sum_{i=1}^{n_{tage}} \eta_i \alpha_{i,j} w_i \right] v_j .\end{aligned} \tag{5.2}$$

As we can see, the first term of Eq. 5.2 is the mean spectrum times its weight obtained from the sum of SFH weight times the SED normalisation factor and the second term is the linear combination of the principal components where the coefficient of each component is weighted by the SFH weight and the normalisation factor. From Eq. 5.2, we can clearly see that we are able to take the advantage of the PCA of the galaxy spectra by replacing the specific fluxes with the eigenvalues.

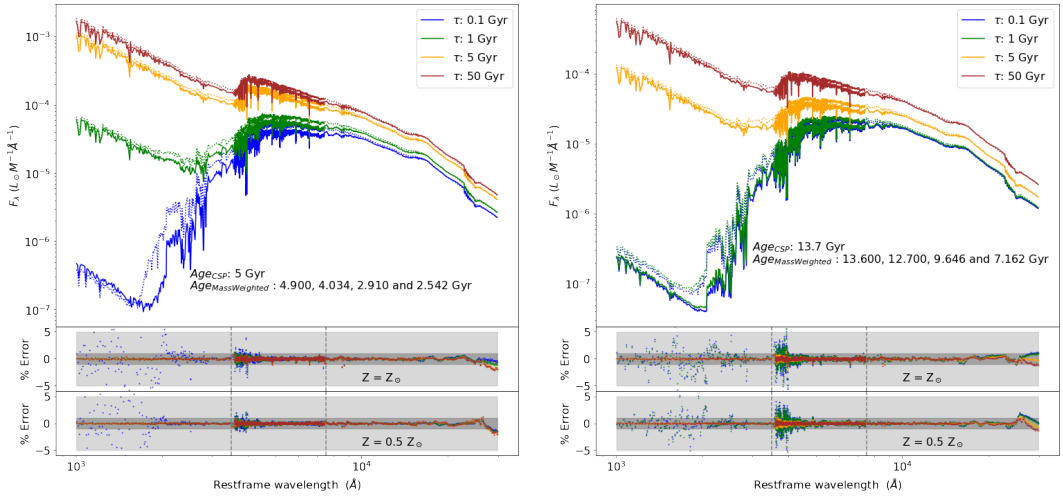
Following this approach, we now show the calculation of composite stellar populations that have different star formation histories, using the τ-model SFH with e-folding times of 0.1, 1.0, 5.0 and 50 Gyr. We also consider two different fixed metallicities, solar and half-solar metallicity. All CSPs are computed at 4 ages including 0.1, 1.0, 5.0 and 13.7 Gyr with 50, 26, and 12 principal components used for the UV, Optical, and IR ranges of the spectrum. The SEDs of these composite populations are shown in Fig. 5.1 and Fig. 5.2. At such ages, the CSPs with different τ value can be distinguished.

In Fig. 5.1 and Fig. 5.2, we calculate the mass-weighted ages of the CSPs and show them as $\mathrm{Age}_{MassWeighted}$ for each value of τ. The mass-weight age gives an indication of the typical age of the SSPs that dominate the composite population. We can clearly see that the composite stellar populations computed using the principal components tend to be more accurate for populations with a small e-folding



a) at 0.1 Gyr b) at 1 Gyr

Figure 5.1: TOP: The CSP SEDs at 0.1 Gyr (a) and 1 (b) Gyr for e-folding times of: 0.1, 1.0, 5.0 and 50 Gyr represented as blue, green, orange and brown lines, respectively. The solar- and half-solar-metallicity CSP are plotted as solid and dotted lines. Note: We only show the reconstructions in the main panel. MIDDLE: The percentage error on the CSP obtained using the PCA-approach compared to the CSP computed using original SSPs as a function of wavelength for solar metallicity. BOTTOM: The same as the middle panel but for the half-solar metallicity CSP.



a) at 5 Gyr b) at 13.7 Gyr

Figure 5.2: The same description as in Fig. 5.1, but now for ages of 5 and 13.7 Gyr.

time, which have less ongoing star-formation at the viewing "age" compared to the CSPs with larger $e-$folding times. A composite population with a short e-folding time is dominated by relatively old SSPs, with a mass-weighted age similar to the age of the galaxy. From the result we obtained in § 4.3.2 our PCA technique fits

the SEDs of very young SSPs better than that of old SSPs, so the CSPs dominated by old SSPs will differ more from the directly computed CSP than the young-age-dominated CSPs, reflecting the relative error in the SSPs. In comparison, we can see that the PCA CSPs tend to fit the spectral features for the half-solar metallicity case better than in the solar metallicity example when considering the CSP at the same age with the same SFH and the same number of principal components used.

To solve these issues, one may perform the PCA on SSPs using a better sampling of the age and metallicity parameter space to improve the SSP reconstruction at high metallicities. In case of the age grid, we could focus on reconstructing the SSP ages that dominate the CSP, i.e. those close to the effective mass-weighted age of the CSP.

5.2 The Photometry of the PCA CSP

A galaxy spectrum provides feature-ful information about a galaxy including the continuum spectrum, absorption lines, emission lines, and spectral breaks (e.g. the Lymann-break and the 4000Å-break). Making use of these spectral features can lead to an understanding of the physical properties of a galaxy. The shift in wavelength of absorption/emission lines is due to the redshift that is directly related its radial velocity (due to the Hubble flow and peculiar velocity). The strengths of some spectral lines can be used a proxy for the morphology of a galaxy. Various properties of model galaxies can be tested against observational data by using the photometry in different bands (e.g. the luminosity function, the Tully-Fisher relation, the colour-morphology relation; see Fig. 2 of Cole et al. 2000). In this section, we will show the calculation of the photometry of the galaxy SEDs in different bands from the UV to the IR. the transmission curves of these bands are shown in Fig.5.3. These filters include the *ugriz* SDSS filters Gunn et al. (1998) and the F115W, F150W, and F200W of the Near Infrared Cammera (NIRCam) of the soon to-be-launched James Webb Space Telescope.

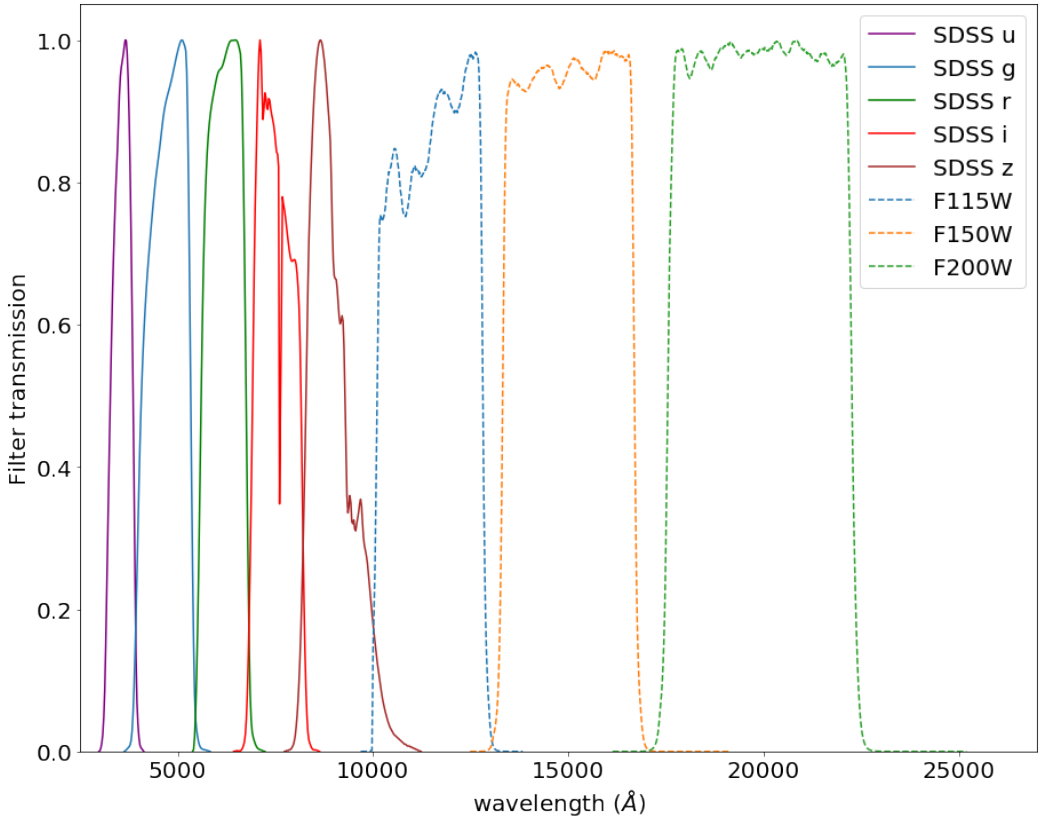


Figure 5.3: The transmission curves of the filters from the Sloan Digital Sky Survey (SDSS) and the James Webb Space Telescope NIRcam instrument. The shape of the curves is the key property for computing the galaxy magnitude. The filters have been normalised to a peak transmission of unity.

We calculate the photometry of the PCA-approached galaxy spectra by using the same numbers of principal components that we used in §5.1. As a result, the magnitudes of the SDSS filters are shown in Table 5.1 and that of the NIRCam is shown in Table 5.2, respectively. The numbers on the top of each element in the tables refer to the percentage errors of magnitudes of the galaxies with solar metallicity whilst the numbers in the bracket at the bottom are for half-solar metallicity CSPs. Overprediction is shown by a positive number. As we can see that, the PCA technique can provide an extraordinary result when we consider the broadband filters. With the total of 88 principal components is able to deliver less than 10^{-2} percent of absolute error for any given age, τ value, and metallicity used in §5.1. This achievement of the PCA technique in calculating the photometry of a galaxy spectrum is what we expect to obtain since the magnitude is the integration

of a galaxy SED through a bandwidth that covers a broad range of wavelength as shown in Fig 5.3. Interestingly we found that a similar level of accuracy can be obtained in the magnitude with only ten principal components.

CSPs		SDSS Filters				
age (Gyr)	τ(Gyr)	u	g	r	i	z
0.1	0.1	6.7E-3 (-1.3E-3)	-1.5E-4 (2.1E-4)	3.5E-4 (7.3E-4)	-5.6E-3 (5.7E-4)	7.7E-4 (-4.5E-4)
	1	5.3E-3 (-1.0E-3)	-2.0E-4 (2.9E-4)	6.0E-4 (5.0E-4)	-6.6E-3 (4.3E-4)	7.2E-4 (-4.7E-4)
	5	5.2E-3 (-1.0E-3)	-2.0E-4 (3.0E-4)	6.2E-4 (4.9E-4)	-6.7E-3 (4.1E-4)	7.2E-4 (-4.7E-4)
	50	5.1E-3 (-1.0E-3)	-2.0E-4 (3.0E-4)	6.2E-4 (4.8E-4)	-6.7E-3 (4.1E-4)	7.2E-4 (-4.7E-4)
1	0.1	-1.7E-2 (-4.4E-3)	4.8E-6 (1.4E-3)	-3.2E-3 (-2.6E-3)	1.6E-3 (7.0E-3)	-7.0E-4 (-1.4E-3)
	1	5.4E-3 (3.0E-3)	-1.5E-4 (6.1E-4)	-1.2E-3 (-6.7E-4)	-2.4E-3 (2.3E-4)	-9.0E-5 (-1.1E-3)
	5	6.0E-3 (-3.1E-3)	-1.6E-4 (5.3E-4)	-8.7E-4 (-4.2E-4)	-3.2E-3 (5.2E-4)	6.0E5 (-9.7E-4)
	50	6.1E-3 (-3.2E-3)	1.6E-4 (5.1E-4)	-8.1E-4 (-3.8E-4)	-3.3E-3 (-5.5E-4)	9.0E-5 (-9.5E-4)
5	0.1	5.6E-3 (4.0E-5)	-6.3E-4 (-9.6E-4)	7.4E-4 (9.2E-4)	2.8E-4 (1.7E-4)	1.9E-4 (9.0E-5)
	1	7.4E-3 (-1.5E-3)	-4.2E-4 (-5.5E-4)	5.5E-4 (6.0E-4)	-1.5E-3 (1.0E-5)	2.7E-4 (-1.4E-4)
	5	-6.1E-3 (-4.1E-3)	-1.6E-4 (2.5E-4)	-2.5E-4 (-1.4E-4)	-3.4E-3 (3.8E-4)	1.5E-4 (-6.3E-4)
	50	6.0E-3 (-4.1E-3)	-1.4E-4 (3.5E-4)	-3.8E-4 (-2.4E-4)	-3.6E-3 (3.8E-4)	1.4E-4 (-7.3E-4)
13.7	0.1	2.0E-2 (3.1E-3)	6.5E-4 (-7.5E-4)	-3.2E-3 (3.3E-4)	6.9E-4 (1.0E-3)	7.0E-5 (3.5E-4)
	1	1.6E-2 (3.2E-3)	4.7E-4 (8.5E-4)	-2.8E-3 (5.7E-4)	1.2E-3 (1.3E-3)	7.0E-5 (3.5E-4)
	5	7.1E-3 (-2.3E-3)	-7.6E-5 (-3.9E-4)	-1.0E-3 (4.8E-4)	-3.8E-4 (1.2E-3)	1.1E-4 (-2.0E-6)
	50	6.3E-3 (-3.8E-3)	-1.5E-4 (9.5E-5)	-5.4E-4 (7.0E-5)	-2.2E-3 (7.8E-4)	1.2E-4 (-4.0E-4)
mean mag error		6.0E-3 (-1.7E-3)	-1.0E-4 (8.5E-5)	-6.6E-4 (4.6E-5)	-2.6E-3 (7.9E-4)	2.1E-4 (-4.6E-4)

Table 5.1: The percentage errors of the PCA CSPs in different SDSS filters compared to the direct CSPs (the numbers inside the bracket for the half-solar metallicity CSPs).

In this study we use a simple parametric form for the star formation history

CSPs		JWST: NIRcam		
Age (Gyr)	τ (Gyr)	f115w	f150w	f200w
0.1	0.1	-8.0E-4 (-1.0E-3)	6.1E-3 (3.2E-3)	-1.1E-2 (-6.5E-3)
	1	-9.8E-4 (-9.8E-4)	6.2E-3 (2.8E-3)	-1.1E-2 (-6.1E-3)
	5	-9.9E-4 (-9.8E-4)	6.2E-3 (2.8E-3)	-1.1E-2 (-6.0E-3)
	50	-1.0E-3 (-9.8E-4)	6.2E-3 (2.8E-3)	-1.1E-2 (-6.0E-3)
1	0.1	-1.1E-3 (-1.8E-3)	7.4E-4 (2.9E-4)	2.3E-3 (4.3E-3)
	1	-8.6E-4 (-8.9E-4)	2.5E-3 (1.4E-3)	-5.0E-4 (4.5E-4)
	5	-8.9E-4 (-8.4E-4)	3.1E-3 (1.7E-3)	-1.9E-3 (-6.1E-4)
	50	-9.0E-4 (-8.3E-4)	3.2E-3 (1.7E-3)	-2.2E-3 (-8.4E-4)
5	0.1	1.3E-3 (5.4E-4)	-4.8E-3 (9.5E-4)	7.6E-3 (-5.0E-3)
	1	9.3E-4 (1.2E-4)	-2.9E-3 (1.5E-3)	5.0E-3 (-5.3E-3)
	5	1.2E-4 (-4.5E-4)	1.1E-4 (1.5E-3)	1.5E-3 (-2.9E-3)
	50	-1.0E-4 (-5.7E-4)	8.9E-4 (1.5E-3)	6.4E-4 (-2.3E-3)
13.7	0.1	3.2E-3 (2.1E-3)	-1.2E-2 (-5.4E-3)	2.5E-2 (4.4E-3)
	1	3.1E-3 (1.9E-3)	-1.2E-3 (-4.7E-3)	2.3E-2 (3.6E-3)
	5	2.0E-3 (8.8E-4)	-6.9E-3 (-1.8E-3)	1.4E-2 (-7.0E-5)
	50	8.3E-4 (7.0E-5)	-2.7E-3 (1.3E-4)	7.7E-3 (-1.5E-3)
mean mag error		2.4E-4 (-2.4E-4)	-3.9E-4 (6.4E-4)	2.4E-3 (-1.9E-3)

Table 5.2: The percentage error of the PCA CSPs in different bands compared to the direct CSPs for JWST NIRCam filters.

called the tau model, and we assume that the beginning of star formation history is also the beginning of the cosmic time. The value of tau (i.e. τ; the e-folding time) we consider can be arbitrary short or very long, for example from 0.01 Gyr to 100 Gyr, which is the recommended range of input value for the tau model for the

FSPS code. However, we can find a realistic range of tau values by comparing the relation between the colour of the PCA galaxy SEDs and their absolute magnitude with an observational colour magnitude diagram (CMD). Here we use the CMD of a sample of galaxies from the Sloan Digital Sky Survey: Data Release 7 (SDSS-DR7 York et al. 2000) where the absolute magnitude in the r-band filter is in a range between -23.5 $< M_r <$ -15.5. To make our reconstructed SEDs comparable with the observational data, we multiply each single SED by $10^{10.5} M_{\odot}$ for the reason that the computed SEDs are normalized to 1 solar mass. Even though the colours of the SDSS:DR7 galaxies are observer frame colours, these galaxies have a low median redshift, $z \approx 0.1$, so we do not attempt to correct the rest-frame for this comparison. Moreover, we roughly adjust the boundary lines for selecting red, green, and blue galaxies from a formula proposed by Papastergis et al. 2013 with a 0.15 mag colour offset to make the lines visually separate the three regions better. The formula we are using is described as the following:

$$g - i = 0.0571(M_r + 24) + C, \tag{5.3}$$

where C is 1.40 for the upper red line and 1.25 for the lower blue line. In contrast, the values of C are 1.25 and 1.10 for the criterion used in Papastergis et al. 2013.

In the comparison between the CMD of the observed galaxies and our model galaxies at the age of 13.7 Gyr as shown in Fig 5.4, we set a range of possible value of τ from 0.01 Gyr to 50 Gyr. By overlapping colour $g - i$ vs. absolute magnitude in r filter of the model SEDs on the CMD, we find that the colour of a galaxy with the value of τ longer than $\sim 10 - 20$ Gyr is extremely "blue" and the value of τ beyond this time scale barely changes the position on the CMD. In summary, it is unnecessary to model a galaxy SED with the e-folding time greater than $\sim 10 - 20$ Gyr for the tau-model SFH since an SED does not show any different in colour for the value excess this which shows relatively high on-going star formation in late time of the history.

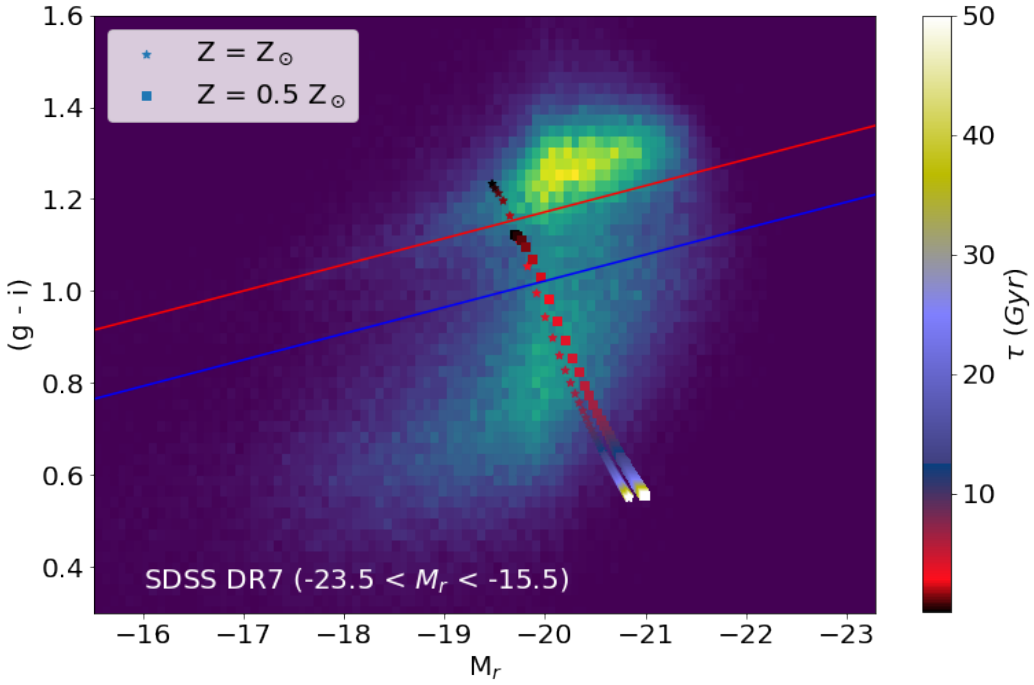


Figure 5.4: The color magnitude diagram of galaxy SEDs obtained from the PCA at an age of 13.7 Gyr, compared with observations. The density plot in the background shows galaxies from SDSS DR7. These are observer frame colours, but SDSS galaxies have a low median redshift, $z \approx 0.1$, so we do not attempt to correct to the rest-frame for this comparison. The red and blue lines represent the boundaries for selecting red, green and blue galaxies in the observations, as proposed by Papastergis et al. (2013), but shifted with a 0.15 mag color offset. Filled symbols show the rest-frame (g-i) colors of the τ models, using the PCA reconstruction, for solar metallicity (stars) and half-solar metallicity (squares). The values of τ used in both cases is indicated by the colormap on the right.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Galaxy spectra are an information-rich tool to study the intrinsic properties of galaxies. Theoretical galaxy formation and evolution models need to be able to predict spectra to allow them to build more realistic mock catalogues. However, the predictions from current galaxy formation models are unlikely to include spectra for all of the model galaxies. A coupling between the model outputs and post-processing methods is needed, for example, by combining a galaxy's star formation history and its chemical evolution with a stellar population synthesis model. In this thesis we focus on an investigation to reduce the computational expense of calculating galaxy spectra by using a data compression method called principal component analysis.

We apply the principal component analysis to a set of the full-wavelength SSP spectra covering 1000 to 30000 Å in 4841 wavelength bins from the FSPS model (Foreman-Mackey et al. 2014). We found that the ability of the PCA to decompose the dimensionality of the data set depends on the data preprocessing method adopted. We studied 4 different preprocessing techniques, including the original spectra, L1-norm spectra, L2-norm spectra, and logarithmic spectra. The logarithmic spectra are the best data set for the PCA to be able to reduce the

dimension of the SSP SEDs. Unfortunately, the logarithmic SSP SEDs can not be used for computing the composite spectra as the CSP SED calculation is a linear combination of each SSP spectrum. Instead, we use result from the PCA of the L2-norm spectra when computing the CSP spectra. Moreover, we found that the SED reconstruction provides a relatively high accuracy for optical spectra whilst providing a poor reconstruction of the UV and NIR spectra when we apply the PCA to the whole wavelength range at once. To reduce the effect of this problem, we compute the principal spectra by dividing each spectrum into 3 separate wavelength ranges; UV (1000-3500Å), Optical (3500-7500 Å), and NIR (7500-30000 Å).

In summary, the PCA can dramatically reduce the dimensionality of the original SSP spectra whilst retaining a relatively high reconstruction accuracy. In Chapter 4 we performed the PCA on simple stellar populations with the metallicity fixed at solar metallicity. This required 68 principal spectra in total (45 for UV, 14 for optical, and 9 for NIR) to reconstruct the SEDs with less than 5% error for the whole of the original FSPS age grid. In addition, 85 principal components in total are needed to rebuild the SEDs with 1% error. These components include 53 UV PCs, 20 optical PCs, and 12 NIR PCs. Moreover, we considered the effects of metallicity on the SSP SEDs by applying the PCA to the simple stellar population SEDs with different metallicities and population ages. As a result, to reconstruct the 2D-parameter-grid SSP SEDs with less than 5% error, one needs 26 optical PCs and 12 NIR PCs. In the case of a target 1% error in the SSP SED reconstruction, 50 optical and 24 NIR components are required. However, we found that the number of UV PCs needed exceeds 100, yet we are not able to rebuild the UV spectra within the two error thresholds stated. By ignoring the same criterion of determining the number of PCs used for the optical and NIR spectra, 50 UV PCs can provide a good fit to the SEDs.

In Chapter 5 the PCA promisingly provides the capability to reduce the computation expense when computing the SED of composite stellar populations. We found that one is able to obtain the composite spectra by using the principal com-

ponents computed in Chapter 4. Moreover, in the case of broadband photometry, the PCA becomes much more effective where the number of components needed decreases dramatically. Fewer than 10 components in total can also yield the magnitude in each filter with similar accuracy as ≈100 components.

6.2 Future work

As presented in the thesis, PCA is a very practical method for decomposing the SSP SEDs which are the building blocks of the galaxy spectra. However, we only considered only basic ingredients of how galaxy spectra are built. The specific parameter space is also limited by the choice of the SPS model we used in this study. To make the application of the PCA to the galaxy spectrum calculation reliable for a galaxy formation model, we could expand upon the work presented here as follows.

- As the ability of PCA to reconstruct the SSP spectra clearly depends on the input sample, the complexity of the SPS models, and the preprocessing technique, an additional effort to find a better sampling of the parameter space and the reprocessing technique could provide a solid improvement when computing the PCA. For example, the parameter space of the input sample is based on the available parameter grids of the SPS model regardless of the SSP SEDs used for the CSP SED calculation where these SSP SEDs are linearly interpolated between the default SPS parameter space. To improve, one may calculate the PCA based on the parameter space of the galaxy formation model (e.g. the star formation history and the chemical evolution history).

- In this thesis, we disregarded the complexity of the SPS model in computing the CSP spectra by only considering the effect of starlight (SSP SED). For example, strong nebular emission lines are a key feature of star-forming

galaxies whilst they are ignored in our calculation. Hence we could make the PCA more realistic by including the effect of nebular emission.

- As the main goal of this study is to reduce the computational expense of generating the full-wavelength-range galaxy spectra for a galaxy formation model. The result shows a promising procedure to solve the problem. Therefore, the PCA-approach spectra calculation then could be implemented into a galaxy formation model when used for computing the galaxy spectra, for example, *GALFORM* (Cole et al. 2000) that provides the star formation history and metallicity as an output.

Appendix **A**

Metallicity Evolution

Although our results show that the fixed-metallicity CSP SEDs are well reproduced using the PCA technique (see §4.3.2), here we consider that the stars that make up CSPs can form with different metallicities. Stars in a galaxy can form at different times from an interstellar medium with an evolving metallicity, governed by the chemical evolution model, which includes the yield of metals from stars and the inflow and outflow of gas (see Cole et al. 2000, Ma et al. 2015). In this section we do not aim to compute CSP SEDs with a realistic metallicity evolution as predicted by a physical model, but instead we want to show that the CSP SED calculation can still be made reliably with a change in metallicity by assuming that the metallicity of stars in the composite population follow a simple linear form described by Equation 6.1.

$$Z(t) = kt + Z_0, \tag{6.1}$$

where Z_0 is the initial stellar metallicity at the beginning of the star formation history and k is the rate of change of metallicity.

Here we consider two examples of composite stellar populations with different star formation histories and different metallicity evolution. The first CSP has its star formation history with $\tau = 1$ Gyr and its metallicity changes from $\log(Z/Z_\odot) = -2.5$ at t = 0 Gyr to $\log(Z/Z_\odot) = 0.5$ at the age of 13.7 Gyr. The second CSP

has $\tau = 5$ Gyr and its final metallicity is $\log(Z/Z_{\odot}) = -1$ with the same initial metallicity. The star formation histories and the metallicity evolutions of these two CSPs are shown in Fig. 6.1. And their corresponding reconstructed SEDs are shown in Fig. 6.2 and Fig. 6.3, respectively.

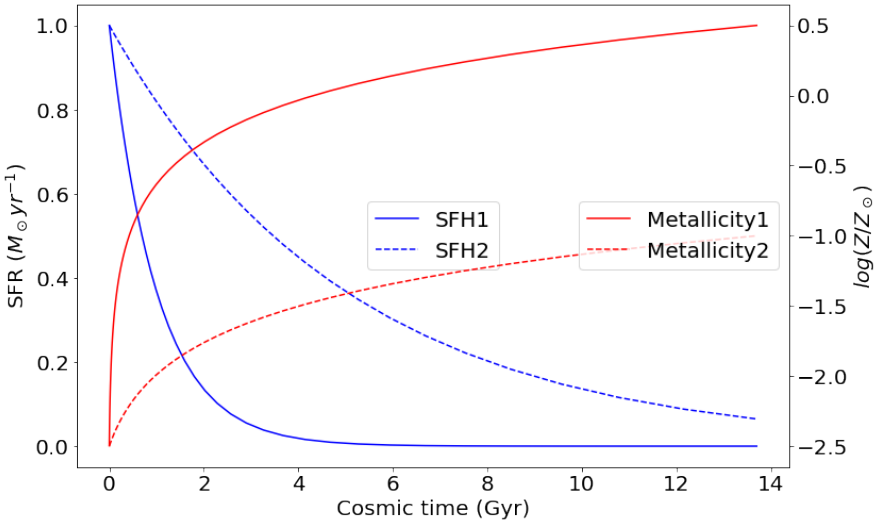


Figure 6.1: The plot shows two τ-model star formation histories with $\tau = 1$ Gyr (blue solid line) and $\tau = 5$ Gyr (blue dashed line) and the adopted metallicity evolution, both starting at $\log(Z/Z_{\odot}) = -2.5$ but one rises up to $\log(Z/Z_{\odot}) = 0.5$ (red solid line) whilst the other reaches $\log(Z/Z_{\odot}) = -1.0$ (red dashed line) at the age of 13.7 Gyr. Note the metallicity is plotted on a logarithmic scale (right hand axis).

In Fig. 6.2 and Fig. 6.3, blue lines represent the original SEDs calculated using the direct output of FSPS model. Purple, green, and red lines in the top panel of each figure show the reconstructed UV, Optical, and NIR spectra and the same colours in the bottom panel show the reconstruction error. We can see that the SEDs of varying-metallicity CSPs are well reconstructed by PCA, using the same number of principal components as we used in the fixed-metallicity CSP calculation.

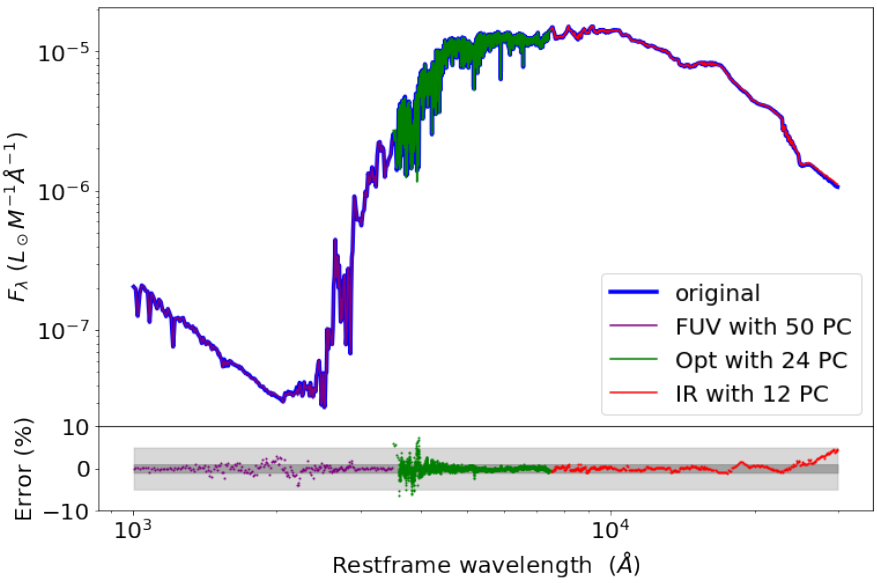


Figure 6.2: The CSP SED at the age of 13.7 Gyr with $\tau = 1$ Gyr and the metallicity changing from $\log(Z/Z_\odot) = -2.5$ to 0.5 associated with the SFH1 and Metallicity1 in Fig. 6.1. The lower panel shows the accuracy of the PCA reconstruction, compared to the direct CSP calculation.

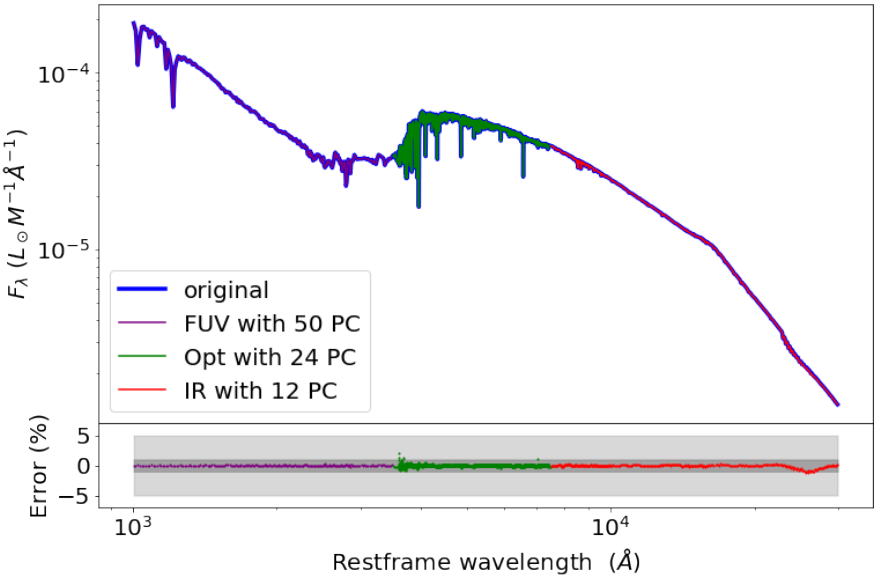


Figure 6.3: The CSP SED at the age of 13.7 Gyr with $\tau = 1$ Gyr and the metallicity changes from $\log(Z/Z_\odot) = -2.5$ to -1.0 associated with the SFH1 and Metallicity1 in Fig. 6.1. The lower panel shows the accuracy of the PCA reconstruction, compared to the direct CSP calculation.

Bibliography

J. Alsing, H. Peiris, J. Leja, C. Hahn, R. Tojeiro, D. Mortlock, B. Leistedt, B. D. Johnson, and C. Conroy. SPECULATOR: Emulating Stellar Population Synthesis for Fast and Accurate Galaxy Spectra and Photometry. *The Astrophysical Journal, Supplement*, 249(1):5, July 2020. doi: 10.3847/1538-4365/ab917f.

C. Baldwin, R. M. McDermid, H. Kuntschner, C. Maraston, and C. Conroy. Comparison of stellar population model predictions using optical and infrared spectroscopy. *Monthly Notices of the Royal Astronomical Society*, 473(4):4698–4721, Feb. 2018. doi: 10.1093/mnras/stx2502.

C. M. Baugh. A primer on hierarchical galaxy formation: the semi-analytical approach. *Reports on Progress in Physics*, 69(12):3101–3156, Dec. 2006. doi: 10.1088/0034-4885/69/12/R02.

G. Bruzual and S. Charlot. Stellar population synthesis at the resolution of 2003. *Monthly Notices of the Royal Astronomical Society*, 344(4):1000–1028, Oct. 2003. doi: 10.1046/j.1365-8711.2003.06897.x.

G. Bruzual A. Spectral evolution of galaxies. I. Early-type systems. *Astrophysical Journal*, 273:105–127, Oct. 1983. doi: 10.1086/161352.

G. Bruzual A. and S. Charlot. Spectral Evolution of Stellar Populations Using

Isochrone Synthesis. *Astrophysical Journal*, 405:538, Mar. 1993. doi: 10.1086/172385.

N. Byler, J. J. Dalcanton, C. Conroy, B. D. Johnson, E. M. Levesque, and D. A. Berg. Stellar and Nebular Diagnostics in the Ultraviolet for Star-forming Galaxies. *Astrophysical Journal*, 863(1):14, Aug. 2018. doi: 10.3847/1538-4357/aacd50.

B. W. Carroll and D. A. Ostlie. *An Introduction to Modern Astrophysics*. Pearson Education, Inc., 1996.

R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966. doi: 10.1207/s15327906mbr0102_10. URL `https://doi.org/10.1207/s15327906mbr0102_10`. PMID: 26828106.

G. Chabrier. Galactic Stellar and Substellar Initial Mass Function. *Publications of the Astronomical Society of the Pacific*, 115(809):763–795, July 2003. doi: 10.1086/376392.

X. Y. Chen, Y. C. Liang, F. Hammer, P. Prugniel, G. H. Zhong, M. Rodrigues, Y. H. Zhao, and H. Flores. Comparing six evolutionary population synthesis models by performing spectral synthesis for galaxies. *Astronomy & Astrophysics*, 515:A101, June 2010. doi: 10.1051/0004-6361/200913894.

Y.-M. Chen, G. Kauffmann, C. A. Tremonti, S. White, T. M. Heckman, K. Kovač, K. Bundy, J. Chisholm, C. Maraston, D. P. Schneider, A. S. Bolton, B. A. Weaver, and J. Brinkmann. Evolution of the most massive galaxies to z= 0.6 - I. A new method for physical parameter estimation. *Monthly Notices of the Royal Astronomical Society*, 421(1):314–332, Mar. 2012. doi: 10.1111/j.1365-2966.2011.20306.x.

J. Choi, A. Dotter, C. Conroy, M. Cantiello, B. Paxton, and B. D. Johnson. Mesa Isochrones and Stellar Tracks (MIST). I. Solar-scaled Models. *Astrophysical Journal*, 823(2):102, June 2016. doi: 10.3847/0004-637X/823/2/102.

S. Cole, C. G. Lacey, C. M. Baugh, and C. S. Frenk. Hierarchical galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 319(1):168–204, Nov. 2000. doi: 10.1046/j.1365-8711.2000.03879.x.

M. Colless, G. Dalton, S. Maddox, W. Sutherland , ..., and K. Taylor. The 2dF Galaxy Redshift Survey: spectra and redshifts. *Monthly Notices of the Royal Astronomical Society*, 328(4):1039–1063, Dec. 2001. doi: 10.1046/j.1365-8711.2001.04902.x.

A. J. Connolly, A. S. Szalay, M. A. Bershady, A. L. Kinney, and D. Calzetti. Spectral Classification of Galaxies: an Orthogonal Approach. *The Astronomical Journal*, 110:1071, Sept. 1995. doi: 10.1086/117587.

C. Conroy. Modeling the Panchromatic Spectral Energy Distributions of Galaxies. *Annual Review of Astronomy & Astrophysics*, 51(1):393–455, Aug. 2013. doi: 10.1146/annurev-astro-082812-141017.

C. Conroy and J. E. Gunn. The Propagation of Uncertainties in Stellar Population Synthesis Modeling. III. Model Calibration, Comparison, and Evaluation. *Astrophysical Journal*, 712(2):833–857, Apr. 2010. doi: 10.1088/0004-637X/712/2/833.

W. I. Cowley, C. M. Baugh, S. Cole, C. S. Frenk, and C. G. Lacey. Predictions for deep galaxy surveys with JWST from ΛCDM. *Monthly Notices of the Royal Astronomical Society*, 474(2):2352–2372, Feb. 2018. doi: 10.1093/mnras/stx2897.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x`.

S. R. Folkes, O. Lahav, and S. J. Maddox. An artificial neural network approach to the classification of galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, 283(2):651–665, Dec. 1996. doi: 10.1093/mnras/283.2.651.

D. Foreman-Mackey, J. Sick, and B. Johnson. Python-Fsps: Python Bindings To Fsps (V0.1.1), Oct. 2014.

A. Gallazzi, S. Charlot, J. Brinchmann, S. D. M. White, and C. A. Tremonti. The ages and metallicities of galaxies in the local universe. *Monthly Notices of the Royal Astronomical Society*, 362(1):41–58, Sept. 2005. doi: 10.1111/j.1365-2966.2005.09321.x.

J. E. Gunn and L. L. Stryker. Stellar spectrophotometric Atlas, 3130 <lambda <10800 A. *The Astrophysical Journal, Supplement*, 52:121–153, June 1983. doi: 10.1086/190861.

J. E. Gunn, M. Carr, C. Rockosi, M. Sekiguchi, ..., and J. Brinkman. The Sloan Digital Sky Survey Photometric Camera. *The Astronomical Journal*, 116(6):3040–3081, Dec. 1998. doi: 10.1086/300645.

I. T. Jolliffe. *Principal Component Analysis and Factor Analysis*, pages 115–128. Springer New York, New York, NY, 1986. ISBN 978-1-4757-1904-8. doi: 10.1007/978-1-4757-1904-8_7. URL `https://doi.org/10.1007/978-1-4757-1904-8_7`.

G. Kauffmann, T. M. Heckman, S. D. M. White, ..., and D. York. Stellar masses and star formation histories for 10^5 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 341(1):33–53, May 2003. doi: 10.1046/j.1365-8711.2003.06291.x.

E. Komatsu, K. M. Smith, J. Dunkley, ..., and E. L. Wright. Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation. *The Astrophysical Journal, Supplement*, 192(2):18, Feb. 2011. doi: 10.1088/0067-0049/192/2/18.

P. Kroupa. On the variation of the initial mass function. *Monthly Notices of the Royal Astronomical Society*, 322(2):231–246, Apr. 2001. doi: 10.1046/j.1365-8711.2001.04022.x.

R. Laureijs, J. Amiaux, S. Arduini, ..., and E. Zucca. Euclid Definition Study Report. *arXiv e-prints*, art. arXiv:1110.3193, Oct. 2011.

C. Leitherer, D. Schaerer, J. D. Goldader, R. M. G. Delgado, C. Robert, D. F. Kune, D. F. de Mello, D. Devost, and T. M. Heckman. Starburst99: Synthesis Models for Galaxies with Active Star Formation. *The Astrophysical Journal, Supplement*, 123(1):3–40, July 1999. doi: 10.1086/313233.

M. E. Levi et al. The Dark Energy Spectroscopic Instrument (DESI). 7 2019.

X. Ma, P. F. Hopkins, C.-A. Faucher-Giguère, N. Zolman, A. L. Muratov, D. Kereš, and E. Quataert. The origin and evolution of the galaxy mass–metallicity relation. *Monthly Notices of the Royal Astronomical Society*, 456(2):2140–2156, 12 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv2659. URL `https://doi.org/10.1093/mnras/stv2659`.

D. S. Madgwick, R. Somerville, O. Lahav, and R. Ellis. Galaxy spectral parametrization in the 2dF Galaxy Redshift Survey as a diagnostic of star formation history. *Monthly Notices of the Royal Astronomical Society*, 343(3):871–879, Aug. 2003. doi: 10.1046/j.1365-8711.2003.06729.x.

C. Maraston. Evolutionary synthesis of stellar populations: a modular tool. *Monthly Notices of the Royal Astronomical Society*, 300(3):872–892, Nov. 1998. doi: 10.1046/j.1365-8711.1998.01947.x.

C. Maraston. Evolutionary population synthesis: models, analysis of the ingredients and application to high-z galaxies. *Monthly Notices of the Royal Astronomical Society*, 362(3):799–825, Sept. 2005. doi: 10.1111/j.1365-2966.2005.09270.x.

C. Maraston, J. Pforr, A. Renzini, E. Daddi, M. Dickinson, A. Cimatti, and C. Tonini. Star formation rates and masses of z $\sim$2 galaxies from multicolour photometry. *Monthly Notices of the Royal Astronomical Society*, 407(2):830–845, Sept. 2010. doi: 10.1111/j.1365-2966.2010.16973.x.

P. D. Mitchell, C. G. Lacey, C. M. Baugh, and S. Cole. How well can we really estimate the stellar masses of galaxies from broad-band photometry? *Monthly Notices of the Royal Astronomical Society*, 435(1):87–114, Oct. 2013. doi: 10.1093/mnras/stt1280.

H. Mo, F. van den Bosch, and S. White. *Galaxy Formation and Evolution*. Cambridge University Press, 2010. doi: 10.1017/CBO9780511807244.

E. Papastergis, R. Giovanelli, M. P. Haynes, A. Rodríguez-Puebla, and M. G. Jones. The Clustering of ALFALFA Galaxies: Dependence on H I Mass, Relationship with Optical Samples, and Clues of Host Halo Properties. *Astrophysical Journal*, 776(1):43, Oct. 2013. doi: 10.1088/0004-637X/776/1/43.

C. Papovich, S. L. Finkelstein, H. C. Ferguson, J. M. Lotz, and M. Giavalisco. The rising star formation histories of distant galaxies and implications for gas accretion with time. *Monthly Notices of the Royal Astronomical Society*, 412(2):1123–1136, Apr. 2011. doi: 10.1111/j.1365-2966.2010.17965.x.

Planck Collaboration, N. Aghanim, Y. Akrami, ..., and A. Zonca. Planck 2018 results. VI. Cosmological parameters. *arXiv e-prints*, art. arXiv:1807.06209, July 2018.

M. J. Rees and J. P. Ostriker. Cooling, dynamics and fragmentation of massive gas clouds: clues to the masses and radii of galaxies and clusters. *Monthly Notices of the Royal Astronomical Society*, 179:541–559, June 1977. doi: 10.1093/mnras/179.4.541.

S. Ronen, A. Aragon-Salamanca, and O. Lahav. Principal component analysis of synthetic galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, 303(2):284–296, Feb. 1999. doi: 10.1046/j.1365-8711.1999.02222.x.

P. Sánchez-Blázquez, R. F. Peletier, J. Jiménez-Vicente, N. Cardiel, A. J. Cenarro, J. Falcón-Barroso, J. Gorgas, S. Selam, and A. Vazdekis. Medium-resolution Isaac Newton Telescope library of empirical spectra. *Monthly No-*

tices of the Royal Astronomical Society, 371(2):703–718, Sept. 2006. doi: 10.1111/j.1365-2966.2006.10699.x.

M. Schmidt. The Rate of Star Formation. *Astrophysical Journal*, 129:243, Mar. 1959. doi: 10.1086/146614.

S. Serneels and T. Verdonck. Principal component analysis for data containing outliers and missing elements. *Computational Statistics & Data Analysis*, 52 (3):1712 – 1727, 2008. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2007.05.024. URL `http://www.sciencedirect.com/science/article/pii/S0167947307002241`.

L. Silva, G. L. Granato, A. Bressan, and L. Danese. Modeling the Effects of Dust on Galactic Spectral Energy Distributions from the Ultraviolet to the Millimeter Band. *Astrophysical Journal*, 509(1):103–117, Dec. 1998. doi: 10.1086/306476.

V. Simha, D. H. Weinberg, C. Conroy, R. Dave, M. Fardal, N. Katz, and B. D. Oppenheimer. Parametrising Star Formation Histories. *arXiv e-prints*, art. arXiv:1404.0402, Apr. 2014.

R. J. Smith. Evidence for initial mass function variation in massive early-type galaxies. *Annual Review of Astronomy and Astrophysics*, 58(1):null, 2020. doi: 10.1146/annurev-astro-032620-020217. URL `https://doi.org/10.1146/annurev-astro-032620-020217`.

R. S. Somerville and R. Davé. Physical Models of Galaxy Formation in a Cosmological Framework. *Annual Review of Astronomy & Astrophysics*, 53:51–113, Aug. 2015. doi: 10.1146/annurev-astro-082812-140951.

B. M. Tinsley. Evolution of the Stars and Gas in Galaxies. *Astrophysical Journal*, 151:547, Feb. 1968. doi: 10.1086/149455.

B. M. Tinsley and J. E. Gunn. Evolutionary synthesis of the stellar population in elliptical galaxies. I. Ingredients, broad-band colors, and infrared features. *Astrophysical Journal*, 203:52–62, Jan. 1976. doi: 10.1086/154046.

J. W. Trayford, P. Camps, T. Theuns, M. Baes, R. G. Bower, R. A. Crain, M. L. P. Gunawardhana, M. Schaller, J. Schaye, and C. S. Frenk. Optical colours and spectral indices of z = 0.1 eagle galaxies with the 3D dust radiative transfer code skirt. *Monthly Notices of the Royal Astronomical Society*, 470(1):771–799, Sept. 2017. doi: 10.1093/mnras/stx1051.

D. H. Weinberg, J. S. Bullock, F. Governato, R. Kuzio de Naray, and A. H. G. Peter. Cold dark matter: Controversies on small scales. *Proceedings of the National Academy of Science*, 112(40):12249–12255, Oct. 2015. doi: 10.1073/pnas.1308716112.

S. D. M. White and C. S. Frenk. Galaxy Formation through Hierarchical Clustering. *Astrophysical Journal*, 379:52, Sept. 1991. doi: 10.1086/170483.

S. D. M. White and M. J. Rees. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. *Monthly Notices of the Royal Astronomical Society*, 183:341–358, May 1978. doi: 10.1093/mnras/183.3.341.

G. Worthey. Comprehensive Stellar Population Models and the Disentanglement of Age and Metallicity Effects. *The Astrophysical Journal, Supplement*, 95:107, Nov. 1994. doi: 10.1086/192096.

R. Yan and MaStar Team. SDSS-IV MaStar: a Large, Comprehensive, and High Quality Empirical Stellar Library. In *Astronomical Society of India Conference Series*, volume 14 of *Astronomical Society of India Conference Series*, pages 99–103, Jan. 2017.

D. G. York, J. Adelman, J. Anderson, John E., ..., N. Yasuda, and SDSS Collaboration. The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120(3):1579–1587, Sept. 2000. doi: 10.1086/301513.

Colophon

This thesis is based on a template developed by Matthew Townson and Andrew Reeves. It was typeset with LaTeX 2_ε. It was created using the *memoir* package, maintained by Lars Madsen, with the *madsen* chapter style. The font used is Latin Modern, derived from fonts designed by Donald E. Kunith.